Решения упражнений

Главы со 2 по 10 | Версия 1.0

Это версия 1.0 руководства по решениям для книги «Глубокое обучение: принципы и концепции» авторов М. Бишопа и Х. Бишопа (Springer, 2024). В нем представлены готовые решения для упражнений в главах 2–10. Полное руководство с решениями всех упражнений будет опубликовано после того, как оно будет получено от авторов книги. Бесплатная цифровая версия решений для покупателей первого издания доступна на веб-сайте книги по адресу `www.dmkpress.com`.

Отзывы о книге или связанных с ней материалах, включая данное руководство с решениями задач, пожалуйста, отправляйте авторам по адресу `dmkpress@gmail.com`.

От переводчика: по причине промежуточного характера этой редакции сборника решений упражнений в ряде фрагментов авторы еще не завершили разметку перекрестных упоминаний различных упражнений, глав книги и других материалов. Эти упоминания отмечены в переводе (как и в оригинале) двумя символами **??** (например, «...аргумент, аналогичный использованному в решении упражнения **??**»).

Глава 2. Вероятности

2.1. Для начала определим $p(T=1)$ путем изменения (2.20):

$$\begin{aligned} p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1) \\ &= \frac{3}{100} \times \frac{999}{1000} + \frac{90}{100} \times \frac{1}{1000} = \frac{3087}{100\,000} = 0,03087. \end{aligned} \tag{1}$$

Затем произведем оценку $p(C=1|T=1)$ путем модификации (2.22):

$$p(C=1|T=1) = \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \tag{2}$$

$$= \frac{90}{100} \times \frac{1}{1000} \times \frac{100\,000}{3087} = \frac{90}{3087} \simeq 0,029. \tag{3}$$

Таким образом, получается, что вероятность заболеть раком даже после положительного теста остается очень малой.

2.2. Необходимо учитывать, что каждое из чисел ряда 0, 1, 2, 3, 4, 5 и 6 выпадает только на одной из этих игральных костей, а это исключает выпадение ничьей при сравнении выпадений одной из костей против другой.

Сначала рассмотрим красную кость и отметим, что на ней имеется четыре копии числа 2 и две копии числа 6. В двух третях случаев при бросании красной кости выпадет 2, и в одной трети случаев выпадет 6. Следовательно, если бросать красную кость в паре с желтой (которая всегда выпадает на 3), желтая кость будет в среднем выигрывать в двух третях случаев, а в одной трети случаев будет проигрывать.

Теперь обратим внимание на синюю кость, на которой четыре раза встречается число 4 и два раза 0. Если бросать ее в паре с желтой костью, она будет выпадать с 4 в двух третях случаев и выигрывать в этом случае и 0 в одной трети случаев и в этом случае будет проигрывать.

Далее рассмотрим зеленую кость в сравнении с синей. Зеленая кость имеет три копии числа 1 и три копии 5. Чтобы вычислить вероятность выигрыша зеленой кости, для начала стоит отметить, что вероятность выпадения 5 на зеленой кости равна 1/2, и в этом случае она обязательно выиграет у синей кости. Точно так же вероятность выпадения 1 на зеленой кости составляет 1/2, и в этом случае вероятность ее выигрыша составляет 1/3. Общая вероятность выигрыша зеленой кости определяется путем умножения вероятностей:

$$\left(\frac{1}{2}\times 1\right)+\left(\frac{1}{2}\times\frac{1}{3}\right)=\frac{2}{3}. \tag{4}$$

Наконец, рассмотрим вероятность выигрыша красной кости в паре с зеленой. Вероятность выпадения числа 6 на красной кости составляет 1/3, и в этом случае она обязательно выигрывает. Точно так же вероятность выпадения числа 2 на красной кости составляет 2/3, и в этом случае вероятность выигрыша этой кости составляет 1/2. Общая вероятность выигрыша красной кости вновь получается путем умножения вероятностей:

$$\left(\frac{1}{3}\times 1\right)+\left(\frac{2}{3}\times\frac{1}{2}\right)=\frac{2}{3}. \tag{5}$$

Для получения дополнительной информации об этих кубиках можно обратиться по адресу `microsoft.com/en-us/research/project/non-transitive-dice/`.

2.3. Желаемое распределение можно выразить в форме с применением правил суммирования и произведения вероятностей:

$$p(\mathbf{y})=\iint p(\mathbf{y}|\mathbf{u},\mathbf{v})p_{\mathbf{u}}(\mathbf{u})p_{\mathbf{v}}(\mathbf{v})\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}. \tag{6}$$

Поскольку $\mathbf{y}$ является детерминированной функцией $\mathbf{u}$ и $\mathbf{v}$, ее условное распределение задается дельта-функцией Дирака в виде

$$p(\mathbf{y}|\mathbf{u}, \mathbf{v}) = \delta(\mathbf{y} - \mathbf{u} - \mathbf{v}). \tag{7}$$

Подстановка (7) в (6) дает возможность выполнить интегрирование по $\mathbf{v}$ и получить

$$p(\mathbf{y}) = \int p_{\mathbf{u}}(\mathbf{u}) p_{\mathbf{v}}(\mathbf{y} - \mathbf{u}) \mathrm{d}\mathbf{u}, \tag{8}$$

как и требовалось.

2.4. Интегрируя равномерное распределение (2.33) по x, получим

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = \int_c^d \frac{1}{d-c}\mathrm{d}x = \frac{d-c}{d-c} = 1. \tag{9}$$

И, следовательно, это распределение является нормализованным. Среднее значение этого распределения составляет

$$\mathbb{E}[x] = \int_a^b \frac{1}{b-a} x\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Дисперсию можно определить, вычислив для начала

$$\mathbb{E}[x^2] = \int_a^b \frac{1}{b-a} x^2\mathrm{d}x = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3},$$

и далее, используя (2.46), получить

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

2.5. Интегрируя экспоненциальное распределение (2.34) в пределах $0 \leq x \leq \infty$, получаем

$$\begin{aligned}\int_0^{\infty} p(x|\lambda)\mathrm{d}x &= \int_0^{\infty} \lambda \exp(-\lambda x)\mathrm{d}x \\ &= \int_0^{\infty} \exp(-y)\mathrm{d}y \\ &= [-\exp(-y)]_0^{\infty} \\ &= 1,\end{aligned} \tag{10}$$

где используется замена переменных $y = \lambda x$. Таким образом, экспоненциальное распределение нормализовано. Точно так же, если проинтегрировать распределение Лапласа (2.35), получится

$$\begin{aligned}\int_{-\infty}^{\infty} p(x|\mu,\lambda)\mathrm{d}x &= \int_{-\infty}^{\infty}\frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)\mathrm{d}x\\ &= \int_{-\infty}^{\mu}\frac{1}{2\gamma}\exp\left(\frac{x-\mu}{\gamma}\right)\mathrm{d}x + \int_{\mu}^{\infty}\frac{1}{2\gamma}\exp\left(-\frac{x-\mu}{\gamma}\right)\mathrm{d}x\\ &= \int_{-\infty}^{0}\frac{1}{2}\exp(z)\mathrm{d}z + \int_{0}^{\infty}\frac{1}{2}\exp(-z)\mathrm{d}z\\ &= \left[\frac{1}{2}\exp(z)\right]_{-\infty}^{0} + \left[-\frac{1}{2}\exp(-z)\right]_{0}^{\infty}\\ &= \frac{1}{2}+\frac{1}{2} = 1,\end{aligned} \tag{11}$$

где в каждом из двух интегралов была произведена подстановка $z = (x - \mu)/\gamma$. Таким образом, получается, что распределение Лапласа также нормализовано.

2.6. Интегрируя эмпирическую плотность вероятности (2.37), получаем

$$\begin{aligned}\int_{-\infty}^{\infty} p(x|\mathcal{D})\mathrm{d}x &= \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(x-x_n)\mathrm{d}x\\ &= \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(y)\mathrm{d}y\\ &= \frac{1}{N}\sum_{n=1}^{N}1 = 1,\end{aligned} \tag{12}$$

как и требовалось. Здесь производится подстановка $y = x - x_n$ в n-й интеграл суммирования, а затем применяется определение дельта-функции Дирака. Этот результат можно легко обобщить на многомерную переменную данных x.

2.7. Если подставить эмпирическое распределение (2.37) в определение математического ожидания по отношению к непрерывной плотности, заданной формулой (2.39), то получится

$$\begin{aligned}\mathbb{E}[f] &= \int_{-\infty}^{\infty} p(x)f(x)\mathrm{d}x\\ &\simeq \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(x-x_n)f(x)\mathrm{d}x\\ &= \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(y_n)f(y_n+x_n)\mathrm{d}y_n\\ &= \frac{1}{N}\sum_{n=1}^{N}f(x_n),\end{aligned} \tag{13}$$

как и требовалось. Здесь применена замена переменной $y_n = x - x_n$ отдельно в каждом из интегралов, а также свойство функции Дирака $\delta(y_n)$, которая интегрируется до единицы и где ненулевой вклад вносит только $y_n = 0$.

2.8. При разложении квадрата получается

$$\begin{aligned}\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2,\end{aligned}$$

как и требовалось.

2.9. Определение ковариации задается формулой (2.47) в виде

$$\mathrm{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Используя (2.38) и тот факт, что $p(x, y) = p(x)p(y)$ при независимых x и y, получаем

$$\begin{aligned}\mathbb{E}[xy] &= \sum_x \sum_y p(x, y)xy \\ &= \sum_x p(x)x \sum_y p(y)y \\ &= \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

и, следовательно, $\mathrm{cov}[x, y] = 0$. Случай, когда x и y являются непрерывными переменными, является аналогичным, при этом (2.38) заменяется на (2.39), а суммы заменяются интегралами.

2.10. Поскольку x и z независимы, их совместное распределение поддается факторизации $p(x, z) = p(x)p(z)$, и поэтому

$$\mathbb{E}[x + z] = \iint (x + z)p(x)p(z)\mathrm{d}x\,\mathrm{d}z \tag{14}$$

$$= \int xp(x)\mathrm{d}x + \int zp(z)\mathrm{d}z \tag{15}$$

$$= \mathbb{E}[x] + \mathbb{E}[z]. \tag{16}$$

Точно так же для дисперсий – сначала следует отметить, что

$$(x + z - \mathbb{E}[x + z])^2 = (x - \mathbb{E}[x])^2 + (z - \mathbb{E}[z])^2 + 2(x - \mathbb{E}[x])(z - \mathbb{E}[z]), \tag{17}$$

где последний член интегрируется по отношению к факторизованному распределению $p(x)p(z)$ до нуля. Таким образом,

$$\begin{aligned}\mathrm{var}[x + z] &= \iint (x + z - \mathbb{E}[x + z])^2 p(x)p(z)\mathrm{d}x\,\mathrm{d}z \\ &= \int (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x + \int (z - \mathbb{E}[z])^2 p(z)\mathrm{d}z \\ &= \mathrm{var}(x) + \mathrm{var}(z).\end{aligned} \tag{18}$$

В случае с дискретными переменными интегралы заменяются суммами, и вновь получаются те же результаты.

2.11. Используя определение математического ожидания (2.39), получаем

$$\begin{aligned}\mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int p(y)\int p(x|y)x\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint p(x,y)x\,\mathrm{d}x\,\mathrm{d}y \\ &= \int p(x)x\,\mathrm{d}x = \mathbb{E}[x], \end{aligned} \tag{19}$$

где применено правило произведения вероятностей $p(x|y)p(y) = p(x, y)$. Теперь используем результат (2.46) для записи

$$\begin{aligned}&\mathbb{E}_y[\mathrm{var}_x[x|y]] + \mathrm{var}_y[\mathbb{E}_x[x|y]] \\ &\quad = \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2. \end{aligned} \tag{20}$$

Теперь отметим, что второй и третий члены в правой части (20) аннулируются. Первый член в правой части (20) можно записать как

$$\begin{aligned}\mathbb{E}_y\mathbb{E}_x[x^2|y] &= \int p(y)\int p(x|y)x^2\mathrm{d}x\,\mathrm{d}y \\ &= \iint p(x,y)x^2\mathrm{d}x\,\mathrm{d}y \\ &= \int p(x)x^2\mathrm{d}x = \mathbb{E}[x^2]. \end{aligned} \tag{21}$$

Таким же образом можно вновь использовать результат (2.46) для записи четвертого члена справа от (20) в виде $\mathbb{E}[x]^2$. Отсюда получаем

$$\mathbb{E}_y[\mathrm{var}_x[x|y]] + \mathrm{var}_y[\mathbb{E}_x[x|y]] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mathrm{var}[x], \tag{22}$$

что и требовалось.

2.12. Преобразование из декартовых координат в полярные определяется выражениями

$$x = r\cos\theta, \tag{23}$$

$$y = r\sin\theta, \tag{24}$$

и, таким образом, получаем $x^2 + y^2 = r^2$, где используется хорошо известное тригонометрическое тождество (3.127). Также видно, что матрица Якоби преобразования переменных равна

$$\begin{aligned}\frac{\partial(x,y)}{\partial(r,\theta)} &= \begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} \end{vmatrix} \\ &= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r, \end{aligned}$$

где вновь используется (3.127). Таким образом, двойной интеграл в (2.125) принимает вид

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\,\mathrm{d}r\,\mathrm{d}\theta \tag{25}$$

$$= 2\pi\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{26}$$

$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)(-2\sigma^2)\right]_0^{\infty} \tag{27}$$

$$= 2\pi\sigma^2, \tag{28}$$

где используется замена переменных $r^2 = u$. Таким образом,

$$I = (2\pi\sigma^2)^{1/2}.$$

Наконец, с помощью преобразования $y = x - \mu$ интеграл гауссова распределения приобретает вид

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x = \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\mathrm{d}y$$
$$= \frac{1}{(2\pi\sigma^2)^{1/2}} = 1,$$

что и требовалось.

2.13. Из определения одномерного гауссова распределения (2.49) получаем

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{(2\pi\sigma^2)}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\mathrm{d}x. \tag{29}$$

Теперь выполним замену переменных $y = x - \mu$ и получим

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{(2\pi\sigma^2)}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}(y+\mu)\mathrm{d}y. \tag{30}$$

Теперь стоит обратить внимание на то, что в множителе $(y + \mu)$ первый член в y соответствует нечетной подынтегральной функции, и по этой причине это интеграл должен исчезнуть (чтобы показать это в явном виде, запишем интеграл как сумму двух интегралов, один от $-\infty$ до 0, а другой от 0 до ∞, а затем наглядно продемонстрируем, что эти два интеграла аннулируются). Во втором слагаемом μ является константой и выносится за пределы интеграла, оставляя нормализованное гауссово распределение, которое, в свою очередь, интегрируется до 1, и, таким образом, получается (2.52).

Чтобы получить (2.53), сначала подставляем выражение (2.49) для нормального распределения в результат нормализации (2.51) и затем перегруппировываем, что дает

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \mathrm{d}x = (2\pi\sigma^2)^{1/2}. \tag{31}$$

Теперь дифференцируем обе стороны (31) относительно σ^2, а затем перегруппировываем, что дает

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 \mathrm{d}x = \sigma^2 \tag{32}$$

и это непосредственно доказывает, что

$$\mathbb{E}[(x-\mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{33}$$

Теперь разложим квадрат в левой части, что дает

$$\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \sigma^2.$$

Используя (2.52), получаем (2.53), как и требовалось.

Наконец, (2.54) следует непосредственно из (2.52) и (2.53):

$$\mathbb{E}[x^2] - \mathbb{E}[x]^2 = (\mu^2 + \sigma^2) - \mu^2 = \sigma^2.$$

2.14. Для одномерного случая достаточно продифференцировать (2.49) по x, чтобы получить

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}(x|\mu,\sigma^2) = -\mathcal{N}(x|\mu,\sigma^2)\frac{x-\mu}{\sigma^2}.$$

Приравнивая это выражение к нулю, получаем $x = \mu$.

2.15. Используем ℓ для обозначения $\ln p(\mathbf{X}|\mu, \sigma^2)$ из (2.56). По стандартным правилам дифференцирования получаем

$$\frac{\mathrm{d}\ell}{\mathrm{d}\mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu).$$

Приравнивая это к нулю и перенося члены, содержащие μ, на другую сторону уравнения, получаем

$$\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n = \frac{1}{\sigma^2} N\mu$$

и далее, умножив обе части на σ^2/N, получаем (2.57).

По аналогии получается

$$\frac{\partial\ell}{\partial\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\frac{1}{\sigma^2},$$

и, приравнивая это к нулю, получаем

$$\frac{N}{2}\frac{1}{\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Умножив обе части на $2(\sigma^2)^2/N$ и подставив μ_{ML} вместо μ, получаем (2.58).

2.16. Если $m = n$, то $x_n x_m = x_n^2$, и при использовании (2.53) получаем $\mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, а если $n \neq m$, то две точки данных x_n и x_m независимы, и, следовательно, $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$, где использовано (2.52). Объединяя эти два результата, получаем (2.128).

Далее получаем

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \mu, \tag{34}$$

где использовано (2.52).

Наконец, рассмотрим $\mathbb{E}[\sigma^2_{\mathrm{ML}}]$. Исходя из (2.57) и (2.58) с использованием (2.128), получаем

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{n=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right] \\
&= \left\{\mu^2 + \sigma^2 - 2\left(\mu^2 + \frac{1}{N}\sigma^2\right) + \mu^2 + \frac{1}{N}\sigma^2\right\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2,
\end{aligned} \tag{35}$$

что и требовалось.

2.17. Из определения (2.61) и с использованием (2.52) и (2.53) получаем

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2 - 2x_n\mu + \mu^2] \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2 - 2\mu\mu + \mu^2) \\
&= \sigma^2,
\end{aligned} \tag{36}$$

как и требовалось.

2.18. Дифференцируя (2.66) по σ^2, получаем

$$\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{2\sigma^4}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2}\frac{1}{\sigma^2}. \tag{37}$$

Приравнивая производную к нулю и производя перегруппировку, получаем

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2, \tag{38}$$

как и требовалось.

2.19. Если сделать предположение, что функция $y = f(x)$ является *строго* монотонной, что необходимо для исключения возможности появления пиков бесконечной плотности в $p(y)$, то тогда будет гарантировано существование обратной функции $x = f^{-1}(y)$. Затем можно использовать (2.71) для записи

$$p(y) = q(f^{-1}(y))\left|\frac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right|. \tag{39}$$

Поскольку единственным ограничением f является ее монотонность, она может произвольно распределять вероятность по x над y. Это показано на рис. 2.12 на стр. 71 как часть решения **??**. Исходя из (39), непосредственно получается, что

$$|f'(x)| = \frac{q(x)}{p(f(x))}.$$

2.20. Матрица Якоби для преобразования из $(x_1; x_2)$ в $(y_1; y_2)$ определяется следующим образом:

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_1}{\partial x_2} \\ \dfrac{\partial y_2}{\partial x_1} & \dfrac{\partial y_2}{\partial x_2} \end{bmatrix}. \tag{40}$$

Для конкретного преобразования, определяемого в (2.78) и (2.79), получается

$$\frac{\partial y_1}{\partial x_1} = 1 + 5\,\mathrm{sech}^2(x_1), \tag{41}$$

$$\frac{\partial y_1}{\partial x_2} = 0, \tag{42}$$

$$\frac{\partial y_2}{\partial x_1} = x_1^2, \tag{43}$$

$$\frac{\partial y_2}{\partial x_2} = 1 + 5\,\mathrm{sech}^2(x_2). \tag{44}$$

2.21. Из описания в вводной части раздела 2.5 следует, что

$$h(p^2) = h(p) + h(p) = 2h(p).$$

Затем допустим, что для всех $k \le K$, $h(p^k) = kh(\mathrm{p})$. Для $k = K + 1$ выполняется

$$h(p^{K+1}) = h(p^K p) = h(p^K) + h(p) = Kh(p) + h(p) = (K + 1)h(p).$$

Кроме того,

$$h(p^{n/m}) = nh(p^{1/m}) = \frac{n}{m} mh(p^{1/m}) = \frac{n}{m} h(p^{m/m}) = \frac{n}{m} h(p),$$

и, следовательно, в соответствии с принципом непрерывности получаем, что $h(p^x) = xh(p)$ для любого действительного числа x.

Теперь рассмотрим положительные действительные числа p и q, а также действительное число x, такие, что $p = q^x$. Из приведенного выше рассуждения следует, что

$$\frac{h(p)}{\ln(p)} = \frac{h(q^x)}{\ln(q^x)} = \frac{x\,h(q)}{x\ln(q)} = \frac{h(q)}{\ln(q)}$$

и, следовательно, $h(p) \propto \ln(p)$.

2.22. Поскольку необходима максимизация энтропии (2.86) при условии, что вероятности в сумме равны единице, получаем

$$\sum_i p(x_i) = 1. \tag{45}$$

Для обеспечения этого ограничения введем множитель Лагранжа λ и, таким образом, получим максимизацию

$$-\sum_i p(x_1)\ln p(x_i) + \lambda\left(\sum_i p(x_1) - 1\right). \tag{46}$$

Устанавливая производную по $p(x_i)$ равной нулю, получаем

$$-\ln p(x_i) - 1 + \lambda = 0. \tag{47}$$

Решая уравнение для $p(x_i)$, получаем

$$p(x_i) = \exp(-1 + \lambda). \tag{48}$$

Поскольку правая часть не зависит от i, это указывает на то, что все вероятности равны. Из (45) следует, что $p(x_i) = 1/M$. Подставляя этот результат в (2.86), получаем, что значение энтропии при ее максимуме равно $\ln M$.

2.23. Энтропия дискретной переменной x с M состояниями может быть записана в виде

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}. \tag{49}$$

Функция $\ln(x)$ является вогнутой $\cap$, поэтому можно применить неравенство Йенсена в виде (2.102), но с обратным неравенством, так что

$$H(x) \leq \ln\left(\sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \right) = \ln M. \tag{50}$$

2.24. Получить требуемую функциональную производную можно путем простого анализа. Если же требуется более формальный подход, можно использовать методы, описанные в приложении В. Рассмотрим сначала функционал

$$I[p(x)] = \int p(x) f(x) \mathrm{d}x.$$

После небольшой модификации $p(x) \to p(x) + \epsilon\eta(x)$ получаем

$$I[p(x) + \epsilon\eta(x)] = \int p(x) f(x) \mathrm{d}x + \epsilon \int \eta(x) f(x) \mathrm{d}x,$$

и, таким образом, из (В.3) следует, что функциональная производная задается формулой

$$\frac{\delta I}{\delta p(x)} = f(x).$$

Аналогичным образом если определить

$$J[p(x)] = \int p(x) \ln p(x) \mathrm{d}x,$$

то при небольшом изменении $p(x) \to p(x) + \epsilon\eta(x)$ получаем

$$J[p(x) + \epsilon\eta(x)] = \int p(x) \ln p(x) \mathrm{d}x + \epsilon \left\{ \int \eta(x) \ln p(x) \mathrm{d}x + \int p(x) \frac{1}{p(x)} \eta(x) \mathrm{d}x \right\} + O(\epsilon^2)$$

и, следовательно,

$$\frac{\delta J}{\delta p(x)} = p(x) + 1.$$

Используя эти два результата, получаем следующий результат для функциональной производной:

$$-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2.$$

После перегруппировки получаем (2.97).

Для устранения множителей Лагранжа подставляем (2.97) в каждое из трех ограничений (2.93), (2.94) и (2.95) по очереди. Решение проще всего получить путем сравнения со стандартной формой гауссовой функции и обращая внимание на то, что результаты

$$\lambda_1 = 1 - \frac{1}{2}\ln(2\pi\sigma^2), \tag{51}$$

$$\lambda_2 = 0, \tag{52}$$

$$\lambda_3 = \frac{1}{2\sigma^2} \tag{53}$$

действительно удовлетворяют трем ограничениям.

Здесь следует отметить, что в вопросе есть опечатка (прим. авторов), и правильно он должен читаться так: «Используйте вариационное исчисление для того, чтобы показать, что стационарная точка функционала, показанного непосредственно перед (1.108), задается в (1.108)».

Многомерную версию этого вывода см. в упражнении 3.8.

2.25. Подставляя правую часть (2.98) в аргумент логарифма правой части (2.91), получаем

$$\begin{aligned}
\mathrm{H}[x] &= -\int p(x)\ln p(x)\mathrm{d}x \\
&= -\int p(x)\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\
&= \frac{1}{2}\left(\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\int p(x)(x-\mu)^2\mathrm{d}x\right) \\
&= \frac{1}{2}(\ln(2\pi\sigma^2) + 1),
\end{aligned}$$

где на последнем шаге используется (2.95).

2.26. Расхождение Кульбака–Лейблера имеет вид

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x})\ln q(\mathbf{x})\mathrm{d}\mathbf{x} + \int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Подставляя функцию гауссова распределения вместо $q(\mathbf{x})$, получаем

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\left\{-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} + \text{const} \\ &= \frac{1}{2}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}]\right)\right\} + \text{const} \\ &= \frac{1}{2}\left\{\ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}] + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]\right)\right\} + \text{const}. \qquad (54)\end{aligned}$$

Дифференцируя это по $\boldsymbol{\mu}$ с использованием результатов из приложения А и приравнивая результат к нулю, получаем

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]. \qquad (55)$$

Точно так же, дифференцируя (54) по $\boldsymbol{\Sigma}^{-1}$ и вновь используя результаты из приложения А, а также (55) и (2.48), получаем

$$\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = \mathrm{cov}[\mathbf{x}]. \qquad (56)$$

2.27. Из (2.100) получаем

$$\mathrm{KL}(p\|q) = -\int p(x)\ln q(x)\mathrm{d}x + \int p(x)\ln p(x)\mathrm{d}x. \qquad (57)$$

Используя (2.49) и (2.51)–(2.53), можно переписать первый интеграл в правой части (57) как

$$\begin{aligned}-\int p(x)\ln q(x)\mathrm{d}x &= \int \mathcal{N}(x|\mu,\sigma^2)\frac{1}{2}\left(\ln(2\pi s^2) + \frac{(x-m)^2}{s^2}\right)\mathrm{d}x \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{1}{s^2}\int \mathcal{N}(x|\mu,\sigma^2)(x^2 - 2xm + m^2)\mathrm{d}x\right) \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2}\right). \qquad (58)\end{aligned}$$

Второй интеграл в правой части (57) получаем из (2.91) как отрицательную дифференциальную энтропию гауссовой функции. Таким образом, из (57), (58) и (2.99) получаем

$$\begin{aligned}\mathrm{KL}(p\|q) &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1 - \ln(2\pi\sigma^2)\right) \\ &= \frac{1}{2}\left(\ln\left(\frac{s^2}{\sigma^2}\right) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1\right).\end{aligned}$$

2.28. Сначала примем $\alpha = 1 - \epsilon$. Тогда

$$p(x)^{(1+\alpha)/2} = p(x)^{1-\epsilon/2} = p(x)\left\{1 - \frac{\epsilon}{2}\ln p(x) + O(\epsilon^2)\right\}. \qquad (59)$$

Точно так же получается, что

$$q(x)^{(1-\alpha)/2} = q(x)^{\epsilon/2} = 1 + \frac{\epsilon}{2}\ln q(x) + O(\epsilon^2). \tag{60}$$

Также имеем

$$1 - \alpha^2 = 1 - 1 + 2\epsilon - \epsilon^2 = 2\epsilon + \mathcal{O}(\epsilon^2). \tag{61}$$

Подставляя эти выражения в альфа-дивергенцию, определенную формулой (2.129), получаем

$$\begin{aligned} \mathrm{D}_\alpha(p\|q) &= \frac{4}{2\epsilon}\left(1 - \int p(x)\left\{1 - \frac{\epsilon}{2}\ln p(x)\right\}\left\{1 + \frac{\epsilon}{2}\ln q(x)\right\}\mathrm{d}x\right) + O(\epsilon) \\ &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x + O(\epsilon), \end{aligned} \tag{62}$$

где используется

$$\int p(x)\mathrm{d}x = 1. \tag{63}$$

Принимая предел $\epsilon \to 0$, получаем $\mathrm{D}_1(p\|q) = \mathrm{KL}(p\|q)$, как и требовалось. Аналогично можно рассмотреть $\alpha = -1 + \epsilon$, что дает

$$p(x)^{(1+\alpha)/2} = p(x)^{\epsilon/2} = 1 + \frac{\epsilon}{2}\ln p(x) + O(\epsilon^2) \tag{64}$$

вместе с

$$q(x)^{(1-\alpha)/2} = q(x)^{1-\epsilon/2} = q(x)\left\{1 - \frac{\epsilon}{2}\ln q(x) + O(\epsilon^2)\right\}. \tag{65}$$

Также имеем

$$1 - \alpha^2 = 1 - 1 + 2\epsilon - \epsilon^2 = 2\epsilon + \mathcal{O}(\epsilon^2). \tag{66}$$

Подставляя эти выражения в альфа-дивергенцию, определенную формулой (2.129), получаем

$$\begin{aligned} \mathrm{D}_\alpha(p\|q) &= \frac{4}{2\epsilon}\left(1 - \int q(x)\left\{1 - \frac{\epsilon}{2}\ln q(x)\right\}\left\{1 + \frac{\epsilon}{2}\ln p(x)\right\}\mathrm{d}x\right) + O(\epsilon) \\ &= -\int q(x)\ln\left\{\frac{p(x)}{q(x)}\right\}\mathrm{d}x + O(\epsilon), \end{aligned} \tag{67}$$

где использовано

$$\int q(x)\mathrm{d}x = 1. \tag{68}$$

Принимая предел $\epsilon \to 0$, получаем $\mathrm{D}_{-1}(p\|q) = \mathrm{KL}(q\|p)$, как и требовалось.

2.29. Сначала используем соотношение $\mathrm{I}(\mathbf{x}; \mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x})$, полученное в (2.110), и отметим, что взаимная информация удовлетворяет условию $\mathrm{I}(\mathbf{x}; \mathbf{y}) \geq 0$, поскольку она является формой дивергенции Кульбака–Лейблера. Затем используем соотношение (2.108) и получим желаемый результат (2.130).

Чтобы показать, что статистическая независимость является достаточным условием для выполнения равенства, подставим $p(\mathbf{x}; \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ в определение энтропии и получим

$$\begin{aligned}\mathrm{H}(\mathbf{x}; \mathbf{y}) &= \iint p(\mathbf{x}; \mathbf{y}) \ln p(\mathbf{x}; \mathbf{y}) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= \iint p(\mathbf{x})p(\mathbf{y})\{\ln p(\mathbf{x}) + \ln p(\mathbf{y})\}\mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= \int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int p(\mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{y} \\ &= \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}).\end{aligned}$$

Чтобы показать, что статистическая независимость является необходимым условием, объединим условие равенства

$$\mathrm{H}(\mathbf{x}; \mathbf{y}) = \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y})$$

с результатом (2.108), что дает

$$\mathrm{H}(\mathbf{y}|\mathbf{x}) = \mathrm{H}(\mathbf{y}).$$

Теперь отметим, что правая часть не зависит от $\mathbf{x}$, а значит, левая часть также должна быть постоянной по отношению к $\mathbf{x}$. Используя (2.110), получаем, что взаимная информация $\mathrm{I}[\mathbf{x}; \mathbf{y}] = 0$. Наконец, используя (2.109), убеждаемся, что взаимная информация является формой расхождения KL, и она исчезает только в том случае, если два распределения равны, так что $p(\mathbf{x}; \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, как и требовалось.

2.30. При замене переменных плотность вероятности преобразуется с помощью матрицы Якоби, определяющей замену переменных. Таким образом, получаем

$$p(\mathbf{x}) = p(\mathbf{y}) \left| \frac{\partial y_i}{\partial x_j} \right| = p(\mathbf{y})|\mathbf{A}|, \tag{69}$$

где $|\cdot|$ обозначает определитель. Тогда энтропия $\mathbf{y}$ может быть записана в виде

$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{y} = -\int p(\mathbf{x})\{\ln p(\mathbf{x})|\mathbf{A}|^{-1}\}\mathrm{d}\mathbf{x} = \mathrm{H}(\mathbf{x}) + \ln|\mathbf{A}|, \tag{70}$$

как и требовалось.

2.31. Условная энтропия $H(y|x)$ может быть записана в виде

$$H(y|x) = -\sum_i \sum_j p(y_i|x_j)p(x_j) \ln p(y_i|x_j), \tag{71}$$

что по условию равно 0. Поскольку величина $-p(y_i|x_j)\ln p(y_i|x_j)$ неотрицательна, каждое из этих слагаемых должно исчезнуть для любого значения x_j, такого, что $p(x_j) \neq 0$. Однако величина $p \ln p$ исчезает только для $p = 0$ или $p = 1$. Таким образом, величины $p(y_i|x_j)$ равны либо 0, либо 1. Однако их сумма должна равняться 1, поскольку это нормализованное распределение вероятностей, и поэтому одна из величин $p(y_i|x_j)$ равна 1, а остальные равны 0. Таким образом, для каждого значения x_j существует единственное значение y_i с ненулевой вероятностью.

2.32. Рассмотрим (2.101) с $\lambda = 0{,}5$ и $b = a + 2\epsilon$ (и, следовательно, $a = b - 2\epsilon$),

$$\begin{aligned} 0{,}5f(a) + 0{,}5f(b) &> f(0{,}5a + 0{,}5b) \\ &= 0{,}5f(0{,}5a + 0{,}5(a + 2\epsilon)) + 0{,}5f(0{,}5(b - 2\epsilon) + 0{,}5b) \\ &= 0{,}5f(a + \epsilon) + 0{,}5f(b - \epsilon). \end{aligned}$$

Это можно переписать как

$$f(b) - f(b - \epsilon) > f(a + \epsilon) - f(a).$$

Затем разделим обе стороны на ϵ и пусть $\epsilon \to 0$, что дает

$$f'(b) > f'(a).$$

Поскольку это верно для всех точек, из этого следует, что $f''(x) \geq 0$ во всех точках.

Чтобы доказать обратное, воспользуемся теоремой Тейлора (с остатком в форме Лагранжа), согласно которой существует такое $x^\star$, что

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x^\star)(x - x_0)^2.$$

Поскольку предполагается, что $f''(x) > 0$ во всех точках, третий член в правой части всегда будет положительным, и поэтому

$$f(x) > f(x_0) + f'(x_0)(x - x_0).$$

Теперь пусть $x_0 = \lambda a + (1 - \lambda)b$ и рассмотрим подстановку $x = a$, что дает

$$\begin{aligned} f(a) &> f(x_0) + f'(x_0)(a - x_0) \\ &= f(x_0) + f'(x_0)((1 - \lambda)(a - b)). \end{aligned} \tag{72}$$

Точно так же при $x = b$ получаем

$$f(b) > f(x_0) + f'(x_0)(\lambda(b - a)). \tag{73}$$

Умножив (72) на λ и (73) на $1 - \lambda$, и затем сложив результаты по обеим сторонам, получаем

$$\lambda f(a) + (1 - \lambda)f(b) > f(x_0) = f(\lambda a + (1 - \lambda)b),$$

как и требовалось.

2.33. Из (2.101) следует, что результат (2.102) справедлив для $M = 1$. Теперь предположим, что он справедлив для некоторого общего значения M, и докажем, что, таким образом, он должен быть справедлив для $M + 1$. Рассмотрим левую часть (2.102):

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) = f\left(\lambda_{M+1}x_{M+1} + \sum_{i=1}^{M}\lambda_i x_i\right) \tag{74}$$

$$= f\left(\lambda_{M+1}x_{M+1} + (1-\lambda_{M+1})\sum_{i=1}^{M}\eta_i x_i\right), \tag{75}$$

где задано

$$\eta_i = \frac{\lambda_i}{1-\lambda_{M+1}}. \tag{76}$$

Теперь применим (2.101), чтобы получить

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \le \lambda_{M+1}f(x_{M+1}) + (1-\lambda_{M+1})f\left(\sum_{i=1}^{M}\eta_i x_i\right). \tag{77}$$

Теперь отметим, что величины λ_i по определению удовлетворяют

$$\sum_{i=1}^{M+1}\lambda_i = 1, \tag{78}$$

и, следовательно, получаем

$$\sum_{i=1}^{M}\lambda_i = 1-\lambda_{M+1}. \tag{79}$$

Затем, используя (76), выясняем, что величины η_i удовлетворяют свойству

$$\sum_{i=1}^{M}\eta_i = \frac{1}{1-\lambda_{M+1}}\sum_{i=1}^{M}\lambda_i = 1. \tag{80}$$

Таким образом, можно применить результат (2.102) для порядка M, и тогда (77) принимает вид

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \le \lambda_{M+1}f(x_{M+1}) + (1-\lambda_{M+1})\sum_{i=1}^{M}\eta_i f(x_i) = \sum_{i=1}^{M+1}\lambda_i f(x_i), \tag{81}$$

где использовано (76).

2.34. В случае с одномерной переменной расхождение KL принимает вид

$$\begin{aligned}\mathrm{KL}(p||q) &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x \\ &= -\int p(x)\ln q(x)\mathrm{d}x + \text{const},\end{aligned} \tag{82}$$

где постоянный член const – это просто отрицательная энтропия фиксированного распределения $p(x)$. Подставляя $p(x)$ в первый член с использованием эмпирического распределения (2.37) и подставляя $q(x)$ с использованием модельного распределения $q(x|\theta)$, получаем

$$\begin{aligned}\mathrm{KL}(p||q) &= -\int \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n)\ln q(x|\theta)\mathrm{d}x + \text{const} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\ln q(x_n|\theta) + \text{const},\end{aligned} \tag{83}$$

что и является искомой отрицательной функцией логарифмического правдоподобия с точностью до аддитивной константы.

2.35. Исходя из (2.92) и используя (2.107), получаем

$$\begin{aligned}\mathrm{H}[\mathbf{x};\mathbf{y}] &= -\iint p(\mathbf{x};\mathbf{y})\ln p(\mathbf{x};\mathbf{y})\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x};\mathbf{y})\ln\left(p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\right)\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x};\mathbf{y})\left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\right)\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x};\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - \iint p(\mathbf{x};\mathbf{y})\ln p(\mathbf{x})\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x};\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - \int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}].\end{aligned}$$

2.36. Для начала оценим маргинальные и условные вероятности $p(x)$, $p(y)$, $p(x|y)$ и $p(y|x)$, чтобы получить результаты, представленные в таблицах ниже. Из всех этих таблиц

x		
	0	2/3
	1	1/3

$p(x)$

	y	
	0	1
	1/3	2/3

$p(y)$

		y	
		0	1
x	0	1	1/2
	1	0	1/2

$p(x|y)$

		y	
		0	1
x	0	1/2	1/2
	1	0	1

$p(y|x)$

и определений

$$H(x) = -\sum_i p(x_i)\ln p(x_i), \tag{84}$$

$$H(x|y) = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j), \tag{85}$$

а также аналогичных определений для $H(y)$ и $H(y|x)$ получаем следующие результаты:

(**a**) $H(x) = \ln 3 - \frac{2}{3}\ln 2$,

(**b**) $H(y) = \ln 3 - \frac{2}{3}\ln 2$,

(**c**) $H(y|x) = \frac{2}{3}\ln 2$,

(**d**) $H(x|y) = \frac{2}{3}\ln 2$,

(**e**) $H(x; y) = \ln 3$,

(**f**) $I(x; y) = \ln 3 - \frac{4}{3}\ln 2$,

где (2.110) используется для оценки взаимной информации. Соответствующая диаграмма показана на рис. 1.

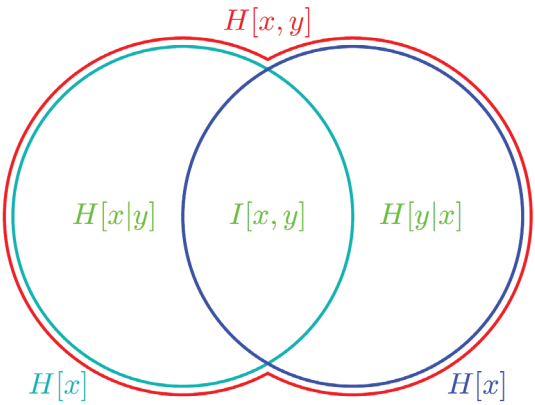

РИС. 1 *Диаграмма взаимосвязей между маргинальной, условной и совместной энтропией и взаимной информацией*

2.37. Арифметическое и геометрическое средние значения определяются как

$$\bar{x}_{\mathrm{A}} = \frac{1}{K}\sum_{k}^{K} x_k \quad \text{и} \quad \bar{x}_{\mathrm{G}} = \left(\prod_{k}^{K} x_k\right)^{1/K}$$

соответственно. Взяв логарифм $\bar{x}_{\mathrm{A}}$ и $\bar{x}_{\mathrm{G}}$, получаем

$$\ln \bar{x}_{\mathrm{A}} = \ln\left(\frac{1}{K}\sum_{k}^{K} x_k\right) \quad \text{и} \quad \ln \bar{x}_{\mathrm{G}} = \frac{1}{K}\sum_{k}^{K} \ln x_k.$$

Сопоставив f с ln, а λ_i с $1/K$ в (2.102) и приняв во внимание, что логарифм является вогнутым, а не выпуклым, в связи с чем неравенство действует в обратном направлении, получаем желаемый результат.

2.38. Из правила произведения вероятностей следует, что $p(\mathbf{x}; \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, и поэтому (2.109) можно записать в виде

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}) &= -\iint p(\mathbf{x}; \mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x}; \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{y} + \iint p(\mathbf{x}; \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{86}$$

2.39. Если z_1 и z_2 независимы, тогда

$$\begin{aligned} \mathrm{cov}[z_1; z_2] &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2) p(z_1; z_2) \mathrm{d}z_1\, \mathrm{d}z_2 \\ &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2) p(z_1) p(z_2) \mathrm{d}z_1\, \mathrm{d}z_2 \\ &= \int (z_1 - \bar{z}_1) p(z_1) \mathrm{d}z_1 \int (z_2 - \bar{z}_2) p(z_2) \mathrm{d}z_2 \\ &= 0, \end{aligned}$$

где

$$\bar{z}_i = \mathbb{E}[z_i] = \int z_i p(z_i) \mathrm{d}z_i.$$

Для y_2 имеем

$$p(y_2|y_1) = \delta(y_2 - y_1^2);$$

т. е. пик вероятностной массы, равный единице, в точке y_1^2, который явно зависит от y_1. Определив $\bar{y}_i$ по аналогии с $\bar{z}_i$ выше, получаем

$$\begin{aligned} \mathrm{cov}[y_1; y_2] &= \iint (y_1 - \bar{y}_1)(y_2 - \bar{y}_2) p(y_1; y_2) \mathrm{d}y_1\, \mathrm{d}y_2 \\ &= \iint y_1 (y_2 - \bar{y}_2) p(y_2|y_1) p(y_1) \mathrm{d}y_1\, \mathrm{d}y_2 \\ &= \int (y_1^3 - y_1 \bar{y}_2) p(y_1) \mathrm{d}y_1 \\ &= 0, \end{aligned}$$

где использован тот факт, что все нечетные моменты y_1 будут равны нулю, поскольку они симметричны относительно нуля.

2.40. В оригинальном издании книги «Глубокое обучение: принципы и концепции» в этом упражнении допущена опечатка: слово «выпуклая» (convex) в первом предложении должно быть заменено на «вогнутая» (concave). Однако обратите внимание, что упражнение можно решить и с использованием слова «выпуклая», следуя тому же алгоритму, что и здесь (прим. авторов). Введем бинарную переменную C, где C = 1 означает, что монета упала вогнутой стороной вверх, и бинарную переменную Q, где Q = 1 означает, что вогнутая сторона монеты является орлом. Исходя из указанных физических

свойств монеты и предполагаемой априорной вероятности того, что вогнутая сторона является орлом, имеем

$$p(\mathrm{C}=1)=0{,}6, \tag{87}$$

$$p(\mathrm{Q}=1)=0{,}1. \tag{88}$$

Набор данных $\mathcal{D}=\{x_1,\ldots,x_{10}\}$ состоит из 10 наблюдений x, каждое из которых принимает значение H (*heads*) для орла или T (*tails*) для решки. Мы предполагаем, что точки данных независимы и идентично распределены, поэтому порядок не имеет значения. Конкретный бросок монеты может привести к выпадению орла либо потому, что она приземлилась вогнутой стороной вверх, а вогнутая сторона является орлом, либо потому, что она приземлилась выпуклой стороной вверх, а вогнутая сторона является решкой. Таким образом, вероятность выпадения орла определяется следующим образом:

$$\begin{aligned} p(x=H) &= p(\mathrm{C}=1)p(\mathrm{Q}=1)+p(\mathrm{C}=0)p(\mathrm{Q}=0)\\ &= 0{,}6\times 0{,}1+0{,}4\times 0{,}9\\ &= 0{,}06+0{,}36\\ &= 0{,}42, \end{aligned} \tag{89}$$

откуда следует, что $p(x=T)=0{,}58$. Вероятность наблюдения 8 орлов и 2 решек, следовательно, равна

$$p(\mathcal{D})=\binom{10}{8}\times(0{,}42)^8\times(0{,}58)^2, \tag{90}$$

где коэффициент в биномиальном разложении задается формулой

$$\binom{N}{K}=\frac{N!}{K!(N-K)!}. \tag{91}$$

Апостериорная вероятность того, что вогнутая сторона будет орлом, определяется по теореме Байеса:

$$p(\mathrm{Q}=1|\mathcal{D})=\frac{p(\mathcal{D}|\mathrm{Q}=1)p(\mathrm{Q}=1)}{p(\mathcal{D})}. \tag{92}$$

Для первого члена в числителе правой части отмечаем, что, так как это обусловлено выпадением вогнутой стороны, нам нужно использовать вероятности выпадения монеты вогнутой стороной вверх и вогнутой стороной вниз, чтобы получить

$$p(\mathcal{D}|\mathrm{Q}=1)=\binom{10}{8}\times(0{,}6)^8\times(0{,}4)^2. \tag{93}$$

Подставляя в теорему Байеса и отмечая, что биномиальные коэффициенты аннулируются, получаем

$$p(\mathrm{Q}=1|\mathcal{D})=\frac{(0,6)^8\times(0,4)^2\times 0,1}{(0,42)^8\times(0,58)^2}\simeq 0,825, \tag{94}$$

и получается, что это более высокая вероятность, чем априорная вероятность 0,1, что интуитивно обосновано, поскольку в наборе данных наблюдалось большее количество орлов по сравнению с решками. Вероятность того, что при следующем подбрасывании выпадет орел, в таком случае определяется как

$$\begin{aligned} p(H|\mathcal{D}) &= p(H|Q=1;\mathcal{D})p(Q=1|\mathcal{D})+p(H|Q=0;\mathcal{D})p(Q=0|\mathcal{D}) \\ &= p(H|Q=1)p(Q=1|\mathcal{D})+p(H|Q=0)p(Q=0|\mathcal{D}) \\ &\simeq 0,6\times 0,825+0,4\times(1-0,825)\simeq 0,565, \end{aligned} \tag{95}$$

что, как видно, выше вероятности 0,42 выпадения орла до наблюдения данных. Опять же, это интуитивно понятно с учетом преобладания орла в наборе данных.

2.41. Если подставить (2.115) в (2.114), получим (2.116). Теперь используем (2.66) для оценки логарифмического правдоподобия линейной регрессионной модели и отметим, что $\ln p(\mathbf{t}|\mathbf{x};\mathbf{w};\sigma^2)$ соответствует $\ln p(\mathcal{D}|\mathbf{w})$. Также отметим, что $\sum_i w_i^2=\mathbf{w}^{\mathrm{T}}\mathbf{w}$. Таким образом, получаем (2.117) для регуляризованной функции ошибки.

Глава 3. Стандартные распределения

3.1. Из определения (3.2) распределения Бернулли получаем

$$\begin{aligned} \sum_{x\in\{0,1\}} p(x|\mu) &= p(x=0|\mu)+p(x=1|\mu) \\ &= (1-\mu)+\mu=1, \end{aligned}$$

$$\sum_{x\in\{0,1\}} xp(x|\mu)=0.p(x=0|\mu)+1.p(x=1|\mu)=\mu,$$

$$\begin{aligned} \sum_{x\in\{0,1\}} (x-\mu)^2p(x|\mu) &= \mu^2p(x=0|\mu)+(1-\mu)^2p(x=1|\mu) \\ &= \mu^2(1-\mu)+(1-\mu)^2\mu=\mu(1-\mu). \end{aligned}$$

Энтропия определяется следующим образом:

$$\begin{aligned} \mathrm{H}[x] &= -\sum_{x\in\{0,1\}} p(x|\mu)\ln p(x|\mu) \\ &= -\sum_{x\in\{0,1\}} \mu^2(1-\mu)^{1-x}\{x\ln\mu+(1-x)\ln(1-\mu)\} \\ &= -(1-\mu)\ln(1-\mu)-\mu\ln\mu. \end{aligned}$$

3.2. Нормализация (3.195) следует из

$$p(x = +1|\mu) + p(x + -1|\mu) = \left(\frac{1+\mu}{2}\right) + \left(\frac{1-\mu}{2}\right) = 1.$$

Среднее значение задается формулой

$$\mathbb{E}[x] = \left(\frac{1+\mu}{2}\right) - \left(\frac{1-\mu}{2}\right) = \mu.$$

Для оценки дисперсии используем

$$\mathbb{E}[x^2] = \left(\frac{1-\mu}{2}\right) + \left(\frac{1+\mu}{2}\right) = 1,$$

откуда получаем

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1 - \mu^2.$$

Наконец, энтропия определяется как

$$\begin{aligned}\mathrm{H}[x] &= -\sum_{x=-1}^{x=+1} p(x|\mu)\ln p(x|\mu) \\ &= -\left(\frac{1-\mu}{2}\right)\ln\left(\frac{1-\mu}{2}\right) - \left(\frac{1+\mu}{2}\right)\ln\left(\frac{1+\mu}{2}\right).\end{aligned}$$

3.3. Используя определение (3.10), получаем

$$\begin{aligned}\binom{N}{n} + \binom{N}{n-1} &= \frac{N!}{n!(N-n)!} + \frac{N!}{(n-1)!(N+1-n)!} \\ &= \frac{(N+1-n)N! + nN!}{n!(N+1-n)!} = \frac{(N+1)!}{n!(N+1-n)!} \\ &= \binom{N+1}{n}. \qquad (96)\end{aligned}$$

Чтобы доказать биномиальную теорему (3.197), обратим внимание, что теорема очевидно верна для $N = 0$. Теперь предположим, что она верна для некоторого общего значения N, и докажем ее истинность для $N + 1$, что можно сделать следующим образом:

$$
\begin{aligned}
(1+x)^{N+1} &= (1+x)\sum_{n=0}^{N}\binom{N}{n}x^n \\
&= \sum_{n=0}^{N}\binom{N}{n}x^n + \sum_{n=1}^{N+1}\binom{N}{n-1}x^n \\
&= \binom{N}{0}x^0 + \sum_{n=1}^{N}\left\{\binom{N}{n}+\binom{N}{n-1}\right\}x^n + \binom{N}{N}x^{N+1} \\
&= \binom{N+1}{0}x^0 + \sum_{n=1}^{N}\binom{N+1}{n}x^n + \binom{N+1}{N+1}x^{N+1} \\
&= \sum_{n=0}^{N+1}\binom{N+1}{n}x^n, \qquad (97)
\end{aligned}
$$

что завершает индуктивное доказательство. Наконец, с помощью биномиальной теоремы условие нормализации (3.198) для биномиального распределения дает

$$
\begin{aligned}
\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} &= (1-\mu)^N\sum_{n=0}^{N}\binom{N}{n}\left(\frac{\mu}{1-\mu}\right)^n \\
&= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N = 1, \qquad (98)
\end{aligned}
$$

как и требовалось.

3.4. Дифференцируя (3.198) по μ, получаем

$$
\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right] = 0.
$$

Умножив на $\mu(1-\mu)$ и произведя перегруппировку, получаем (3.11).

Если дважды дифференцировать (3.198) по μ, то получим

$$
\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left\{\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right]^2-\frac{n}{\mu^2}-\frac{(N-n)}{(1-\mu)^2}\right\} = 0.
$$

Теперь умножим на $\mu^2(1-\mu)^2$ и перегруппируем, используя результат (3.11) для среднего значения биномиального распределения, чтобы получить

$$
\mathbb{E}[n^2] = N\mu(1-\mu)+N^2\mu^2.
$$

Наконец, используем (2.46), чтобы получить результат (3.12) для дисперсии.

3.5. Дифференцируем (3.26) по $\mathbf{x}$, чтобы получить

$$\begin{aligned}\frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\nabla_{\mathbf{x}}\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \\ &= -\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),\end{aligned}$$

где используем (A.19), (A.20) и тот факт, что $\boldsymbol{\Sigma}^{-1}$ симметрична. Приравнивая эту производную к 0 и умножая слева на $\boldsymbol{\Sigma}$, получаем решение $\mathbf{x} = \boldsymbol{\mu}$.

3.6. Сначала определим $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ и предположим, что $\mathbf{y}$ имеет ту же размерность, что и $\mathbf{x}$, так что $\mathbf{A}$ является квадратной матрицей. Также предположим, что матрица $\mathbf{A}$ является симметричной и имеет обратную матрицу. Отсюда следует, что $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$. Из правил суммирования и умножения вероятностей получаем

$$\begin{aligned}p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int \delta(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})p(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\propto \int \delta(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))p(\mathbf{x})\mathrm{d}\mathbf{x}, \end{aligned} \tag{99}$$

где $\delta(\cdot)$ – это дельта-функция Дирака. Подставляя $p(\mathbf{x})$ с использованием гауссова распределения, получаем

$$\begin{aligned}p(\mathbf{y}) &\propto \int \delta(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\mathbf{x} \\ &\propto \int \delta(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) - \boldsymbol{\mu})\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{b} - \mathbf{A}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b} - \mathbf{A}\boldsymbol{\mu})\right\}, \end{aligned} \tag{100}$$

где используется свойство симметричной матрицы, согласно которому ее обратная матрица также является симметричной. При анализе видно, что $p(\mathbf{y})$ является гауссовым распределением, его среднее значение равно $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$, а ковариация имеет вид

$$(\mathbf{A}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1})^{-1} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}. \tag{101}$$

3.7. Из (2.100) получаем

$$\mathrm{KL}(q(\mathbf{x})\|p(\mathbf{x})) = -\int q(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x} + \int q(\mathbf{x})\ln q(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{102}$$

Используя (3.26), (3.40), (3.42) и (3.46), можно переписать первый интеграл в правой части (102) как

$$
\begin{aligned}
&-\int q(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\quad = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)\frac{1}{2}\{D\ln(2\pi) + \ln|\boldsymbol{\Sigma}_p| + (\mathbf{x}-\boldsymbol{\mu}_p)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}(\mathbf{x}-\boldsymbol{\mu}_p)\}\mathrm{d}\mathbf{x} \\
&\quad = \frac{1}{2}\{D\ln(2\pi) + \ln|\boldsymbol{\Sigma}_p| + \mathrm{Tr}[\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q\boldsymbol{\mu}_q^{\mathrm{T}} + \boldsymbol{\Sigma}_q)] \\
&\qquad - \boldsymbol{\mu}_p\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_q - \boldsymbol{\mu}_q^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p + \boldsymbol{\mu}_q^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p\}. \qquad (103)
\end{aligned}
$$

Второй интеграл в правой части (102) выясняется из (2.92) в качестве отрицательной дифференциальной энтропии многомерного гауссова распределения. Таким образом, из (102), (103) и (3.204) получаем

$$
\mathrm{KL}(p(\mathbf{x})\|p(\mathbf{x})) = \frac{1}{2}\left\{\ln\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - D + \mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right\}. \qquad (104)
$$

3.8. Можно использовать множители Лагранжа для обеспечения ограничений для решения с максимальной энтропией. Следует отметить, что для ограничения нормализации (3.201) требуется один множитель Лагранжа, для D ограничений, заданных формулой (3.202), требуется D-мерный вектор $\mathbf{m}$ множителей Лагранжа, а для обеспечения D^2 ограничений, представленных формулой (3.203), требуется $D\times D$-матрица $\mathbf{L}$ множителей Лагранжа. Таким образом, максимизируем

$$
\begin{aligned}
\tilde{\mathrm{H}}[p] = &-\int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x} + \lambda\Big(\int p(\mathbf{x})\mathrm{d}\mathbf{x} - 1\Big) \\
&+ \mathbf{m}^{\mathrm{T}}\Big(\int p(\mathbf{x})\mathbf{x}\mathrm{d}\mathbf{x} - \boldsymbol{\mu}\Big) \\
&+ \mathrm{Tr}\Big\{\mathbf{L}\Big(\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{d}\mathbf{x} - \boldsymbol{\Sigma}\Big)\Big\}. \qquad (105)
\end{aligned}
$$

По функциональному дифференцированию (приложение B) максимум этого функционала по $p(\mathbf{x})$ достигается при

$$
0 = -1 - \ln p(\mathbf{x}) + \lambda + \mathbf{m}^{\mathrm{T}}\mathbf{x} + \mathrm{Tr}\{\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\}.
$$

Решая уравнение для $p(\mathbf{x})$, получаем

$$
p(\mathbf{x}) = \exp\{\lambda - 1 + \mathbf{m}^{\mathrm{T}}\mathbf{x} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})\}. \qquad (106)
$$

Теперь находим значения множителей Лагранжа с применением ограничений. Сначала доводим квадрат внутри экспоненты до полного квадрата, который принимает вид

$$
\lambda - 1 + \left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}\right)^{\mathrm{T}}\mathbf{L}\left(\mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}\right) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \frac{1}{4}\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m}.
$$

Теперь производим замену переменной:

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}.$$

Тогда ограничение (3.202) принимает вид

$$\int \exp\left\{\lambda - 1 + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \frac{1}{4}\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m}\right\}\left(\mathbf{y} + \boldsymbol{\mu} - \frac{1}{2}\mathbf{L}^{-1}\mathbf{m}\right)\mathrm{d}\mathbf{y} = \boldsymbol{\mu}.$$

В последних скобках член в $\mathbf{y}$ исчезает по симметрии, а член в $\boldsymbol{\mu}$ просто интегрируется в $\boldsymbol{\mu}$ благодаря ограничению нормализации (3.201), которое теперь принимает вид

$$\int \exp\left\{\lambda - 1 + \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{m} - \frac{1}{4}\mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m}\right\}\mathrm{d}\mathbf{y} = 1,$$

и, следовательно, получаем

$$-\frac{1}{2}\mathbf{L}^{-1}\mathbf{m} = 0,$$

где снова использовано ограничение (3.201). Таким образом, $\mathbf{m} = \mathbf{0}$, и плотность принимает вид

$$p(\mathbf{x}) = \exp\{\lambda - 1 + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x} - \boldsymbol{\mu})\}.$$

Подставляя это в окончательное ограничение (3.203) и производя замену переменной $\mathbf{x} - \boldsymbol{\mu} = \mathbf{z}$, получаем

$$\int \exp\{\lambda - 1 + \mathbf{z}^{\mathrm{T}}\mathbf{L}\mathbf{z}\}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathrm{d}\mathbf{x} = \boldsymbol{\Sigma}.$$

Применив аналогичный аргумент, который использовался для вывода (3.48), получаем $\mathbf{L} = -\frac{1}{2}\boldsymbol{\Sigma}$. Наконец, значение λ просто является тем значением, которое необходимо для обеспечения правильной нормализации гауссова распределения, как выведено в разделе 3.2, и, следовательно, задается формулой

$$\lambda - 1 = \ln\left\{\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\right\}.$$

3.9. Из определений многомерной дифференциальной энтропии (2.92) и многомерного гауссова распределения (3.26) получаем

$$\begin{aligned}\mathrm{H}[\mathbf{x}] &= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \mathrm{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\mathbf{x} \\ &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\frac{1}{2}(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))\mathrm{d}\mathbf{x}\end{aligned}$$

$$= \frac{1}{2}(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}])$$
$$= \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + \ln(2\pi)).$$

3.10. Нам известно, что $p(x_1) = \mathcal{N}(x_1|\mu_1;\ \tau_1^{-1})$ и $p(x_2) = \mathcal{N}(x_2|\mu_2;\ \tau_2^{-1})$. Поскольку $x = x_1 + x_2$, также известно, что $p(x|x_2) = \mathcal{N}(x|\mu_1 + x_2;\ \tau_1^{-1})$. Теперь вычислим интеграл свертки, заданный формулой (3.205), который имеет вид

$$p(x) = \left(\frac{\tau_1}{2\pi}\right)^{1/2}\left(\frac{\tau_2}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2 - \frac{\tau_2}{2}(x_2 - \mu_2)^2\right\}\mathrm{d}x_2. \qquad (107)$$

Поскольку конечным результатом будет гауссово распределение для $p(x)$, нам достаточно оценить только его точность, так как из (2.99) следует, что энтропия определяется дисперсией или, что эквивалентно, точностью и не зависит от среднего значения. Это позволяет упростить вычисления и игнорировать такие факторы, как константы нормализации.

Начнем с рассмотрения слагаемых в экспоненте (107), которые зависят от x_2 и задаются формулой

$$-\frac{1}{2}x_2^2(\tau_1 + \tau_2) + x_2\{\tau_1(x - \mu_1) + \tau_2\mu_2\}$$
$$= -\frac{1}{2}(\tau_1 + \tau_2)\left\{x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right\}^2 + \frac{\{\tau_1(x - \mu_1) + \tau_2\mu_2\}^2}{2(\tau_1 + \tau_2)},$$

где производится полное квадратное завершение по x_2. При интегрировании x_2 первый член в правой части даст простую константу, не зависящую от x. Второй член при разложении будет включать член в x^2. Поскольку точность x задается непосредственно коэффициентом x^2 в показателе степени, нам необходимо учитывать только такие члены. Есть еще один член в x^2, возникающий из исходного показателя в (107). Объединяя их, получаем

$$-\frac{\tau_1}{2}x^2 + \frac{\tau_1^2}{2(\tau_1 + \tau_2)}x^2 = -\frac{1}{2}\frac{\tau_1\tau_2}{\tau_1 + \tau_2}x^2,$$

из чего видно, что x имеет точность $\tau_1\tau_2/(\tau_1 + \tau_2)$.

Этот результат для точности также можно получить непосредственно, обратившись к общему результату (3.99) для свертки двух линейных гауссовых распределений.

Энтропия x в этом случае определяется из (2.99) как

$$\mathrm{H}[x] = \frac{1}{2}\ln\left\{\frac{2\pi(\tau_1 + \tau_2)}{\tau_1\tau_2}\right\}.$$

3.11. Мы можем использовать аргумент, аналогичный использованному в решении упражнения **??**. Рассмотрим общую квадратную матрицу $\mathbf{\Lambda}$ с элементами Λ_{ij}. Тогда всегда можно записать $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$, где

$$\Lambda_{ij}^{\mathrm{S}} = \frac{\Lambda_{ij} + \Lambda_{ji}}{2}, \qquad \Lambda_{ij}^{\mathrm{A}} = \frac{\Lambda_{ij} - \Lambda_{ji}}{2}, \tag{108}$$

и удостовериться, что $\mathbf{\Lambda}^{\mathrm{S}}$ является симметричной, так что $\mathbf{\Lambda}_{ij}^{\mathrm{S}} = \mathbf{\Lambda}_{ji}^{\mathrm{S}}$, а $\mathbf{\Lambda}^{\mathrm{A}}$ является антисимметричной, так что $\mathbf{\Lambda}_{ij}^{\mathrm{A}} = -\mathbf{\Lambda}_{ji}^{\mathrm{S}}$. Квадратичная форма в показателе степени D-мерного многомерного гауссова распределения может быть записана в виде

$$\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i - \mu_i)\Lambda_{ij}(x_j - \mu_j), \tag{109}$$

где $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ – это матрица точности. После подстановки $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ в (109) получается, что член, содержащий $\mathbf{\Lambda}^{\mathrm{A}}$, исчезает, поскольку для каждого положительного члена существует равный и противоположный отрицательный член. Таким образом, $\mathbf{\Lambda}$ можно всегда считать симметричной.

3.12. Начнем с предварительного умножения обеих частей (3.28) на $\mathbf{u}_i^{\dagger}$, сопряженную транспонированную матрицу $\mathbf{u}_i$. Это дает

$$\mathbf{u}_i^{\dagger}\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i^{\dagger}\mathbf{u}_i. \tag{110}$$

Далее рассмотрим сопряженную транспонированную матрицу (3.28) и умножим ее на $\mathbf{u}_i$, что дает

$$\mathbf{u}_i^{\dagger}\mathbf{\Sigma}^{\dagger}\mathbf{u}_i = \lambda_i^{*}\mathbf{u}_i^{\dagger}\mathbf{u}_i, \tag{111}$$

где λ_i^{*} – это комплексное сопряжение λ_i. Теперь вычитаем (110) из (111) и используем тот факт, что $\mathbf{\Sigma}$ является действительной и симметричной, и, следовательно, $\mathbf{\Sigma} = \mathbf{\Sigma}^{\dagger}$, чтобы получить

$$0 = (\lambda_i^{*} - \lambda_i)\mathbf{u}_i^{\dagger}\mathbf{u}_i.$$

Следовательно, $\lambda_i^{*} = \lambda_i$, а значит, λ_i должны быть действительными.
Теперь рассмотрим

$$\begin{aligned}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\lambda_j &= \mathbf{u}_i^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u}_j \\ &= \mathbf{u}_i^{\mathrm{T}}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{u}_j \\ &= (\mathbf{\Sigma}\mathbf{u}_i)^{\mathrm{T}}\mathbf{u}_j \\ &= \lambda_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j,\end{aligned}$$

где использовано (3.28) и тот факт, что $\mathbf{\Sigma}$ является симметричной. Если предположить, что $0 \neq \lambda_i \neq \lambda_j \neq 0$, единственным решением этого уравнения будет $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$, т. е. $\mathbf{u}_i$ и $\mathbf{u}_j$ ортогональны.

Если $0 \neq \lambda_i = \lambda_j \neq 0$, любая линейная комбинация $\mathbf{u}_i$ и $\mathbf{u}_j$ будет собственным вектором с собственным значением $\lambda = \lambda_i = \lambda_j$, поскольку из (3.28) следует, что

$$\begin{aligned}\boldsymbol{\Sigma}(a\mathbf{u}_i + b\mathbf{u}_j) &= a\lambda_i\mathbf{u}i + b\lambda_j\mathbf{u}_j \\ &= \lambda(a\mathbf{u}_i + b\mathbf{u}_j).\end{aligned}$$

Исходя из того, что $\mathbf{u}_i \neq \mathbf{u}_j$, можно составить

$$\mathbf{u}_\alpha = a\mathbf{u}_i + b\mathbf{u}_j,$$

$$\mathbf{u}_\beta = c\mathbf{u}_i + d\mathbf{u}_j$$

таким образом, что $\mathbf{u}_\alpha$ и $\mathbf{u}_\beta$ взаимно ортогональны и имеют единичную длину. Поскольку $\mathbf{u}_i$ и $\mathbf{u}_j$ ортогональны $\mathbf{u}_k$ ($k \neq i, k \neq j$), то $\mathbf{u}_\alpha$ и $\mathbf{u}_\beta$ также ортогональны. Таким образом, $\mathbf{u}_\alpha$ и $\mathbf{u}_\beta$ удовлетворяют (3.29).

Наконец, если $\lambda_i = 0$, $\boldsymbol{\Sigma}$ должна быть сингулярной, при этом $\mathbf{u}_i$ лежит в нулевом пространстве $\boldsymbol{\Sigma}$. В этом случае $\mathbf{u}_i$ будет ортогонален собственным векторам, которые проецируются на пространство строк $\boldsymbol{\Sigma}$, и теперь можно выбрать $\|\mathbf{u}_i\| = 1$, чтобы это удовлетворяло (3.29). Если более одного собственного значения равно нулю, можно выбрать соответствующие собственные векторы произвольно, при условии, что они остаются в нулевом пространстве $\boldsymbol{\Sigma}$, и, таким образом, можно выбрать их так, чтобы они удовлетворяли (3.29).

3.13. Правую часть (3.31) можно записать в матричной форме как

$$\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{M},$$

где $\mathbf{U}$ – это матрица $D \times D$, столбцами которой являются собственные векторы $\mathbf{u}_1, \ldots, \mathbf{u}_D$, а $\boldsymbol{\Lambda}$ – это диагональная матрица, на диагонали которой находятся собственные значения $\lambda_1, \ldots, \lambda_D$.

Таким образом, получается

$$\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\mathbf{U} = \boldsymbol{\Lambda}.$$

Тем не менее из (3.28)–(3.30) также получается, что

$$\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda} = \boldsymbol{\Lambda},$$

и, следовательно, $\mathbf{M} = \boldsymbol{\Sigma}$ и (3.31) выполняется.

Кроме того, поскольку $\mathbf{U}$ является ортонормированной, $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$, и, следовательно,

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}})^{-1} = (\mathbf{U}^{\mathrm{T}})^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}.$$

3.14. Поскольку $\mathbf{u}_1, \ldots, \mathbf{u}_D$ составляют базис для $\mathbb{R}^D$, можно записать

$$\mathbf{a} = \hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \ldots + \hat{a}_D\mathbf{u}_D,$$

где $\hat{a}_1, \ldots, \hat{a}_D$ – это коэффициенты, полученные путем проецирования $\mathbf{a}$ на $\mathbf{u}_1, \ldots, \mathbf{u}_D$. Здесь следует отметить, что они, как правило, *не равны* элементам $\mathbf{a}$.

Используя это, можно записать

$$\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a} = (\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}})\boldsymbol{\Sigma}(\hat{a}_1\mathbf{u}_1 + \ldots + \hat{a}_D\mathbf{u}_D),$$

и, объединив этот результат с (3.28), получаем

$$(\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}})(\hat{a}_1\lambda_1\mathbf{u}_1 + \ldots + \hat{a}_D\lambda_D\mathbf{u}_D).$$

Теперь, поскольку $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 1$, только если $i = j$, а в противном случае 0, это преобразуется в

$$\hat{a}_1^2\lambda_1 + \ldots + \hat{a}_D^2\lambda_D,$$

и поскольку $\mathbf{a}$ является действительным, получается, что это выражение будет строго положительным для любого ненулевого $\mathbf{a}$, если все собственные значения строго положительны. Также очевидно, что если собственное значение λ_i равно нулю или отрицательно, то существует вектор $\mathbf{a}$ (например, $\mathbf{a} = \mathbf{u}_i$), для которого это выражение будет меньше или равно нулю. Таким образом, наличие у матрицы собственных векторов, которые все строго положительны, является достаточным и необходимым условием для того, чтобы матрица была положительно определенной.

3.15. Матрица $D \times D$ имеет D^2 элементов. Если она симметрична, то элементы, не лежащие на главной диагонали, образуют пары равных значений. На диагонали находится D элементов, поэтому количество элементов, не лежащих на диагонали, равно $D^2 - D$, и только половина из них являются независимыми, что дает

$$\frac{D^2 - D}{2}.$$

Если теперь добавить обратно D элементов на диагонали, то получится

$$\frac{D^2 - D}{2} + D = \frac{D(D+1)}{2}.$$

3.16. Рассмотрим матрицу $\mathbf{M}$, которая является симметричной, так что $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$. Обратная матрица $\mathbf{M}^{-1}$ выполняет условие

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}.$$

Взяв транспонированную матрицу обеих частей этого уравнения и используя соотношение (A.1), получаем

$$(\mathbf{M}^{-1})^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I},$$

поскольку единичная матрица является симметричной. Используя условие симметричности для $\mathbf{M}$, получаем

$$(\mathbf{M}^{-1})^{\mathrm{T}}\mathbf{M} = \mathbf{I},$$

и, следовательно, из определения обратной матрицы

$$(\mathbf{M}^{-1})^{\mathrm{T}} = \mathbf{M}^{-1},$$

а значит, $\mathbf{M}^{-1}$ также является симметричной матрицей.

3.17. Вспомним, что преобразование (3.34) диагонализирует систему координат и что квадратичная форма (3.27), соответствующая квадрату расстояния Махаланобиса, в этом случае задается формулой (3.33). Это соответствует сдвигу начала координат и повороту, в результате чего гиперэллипсоидальные контуры, вдоль которых расстояние Махаланобиса остается постоянным, становятся осевыми. Объем, содержащийся в любом таком контуре, не изменяется при сдвигах и поворотах. Теперь производим дальнейшее преобразование $z_i = \lambda_i^{1/2} y_i$ для $i = 1, \ldots, D$. Объем внутри гиперэллипсоида в этом случае будет равен

$$\int \prod_{i=1}^{D} \mathrm{d}y_i = \prod_{i=1}^{D} \lambda_i^{1/2} \int \prod_{i=1}^{D} \mathrm{d}z_i = |\mathbf{\Sigma}|^{1/2} V_D \Delta^D,$$

где использовано свойство, что определитель $\mathbf{\Sigma}$ задается произведением его собственных значений, а также тот факт, что в координатах z объем стал сферой радиуса Δ, объем которой равен $V_D \Delta^D$.

3.18. Умножение левой части (3.60) на матрицу (3.208) дает тождественную матрицу. В правой части рассмотрим четыре блока полученной разделенной матрицы:

вверху слева

$$\mathbf{AM} - \mathbf{BD}^{-1}\mathbf{CM} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} = \mathbf{I}$$

вверху справа

$$\begin{aligned} &-\mathbf{AMBD}^{-1} + \mathbf{BD}^{-1} + \mathbf{BD}^{-1}\mathbf{CMBD}^{-1} \\ &\quad = -(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} + \mathbf{BD}^{-1} \\ &\quad = -\mathbf{BD}^{-1} + \mathbf{BD}^{-1} = \mathbf{0} \end{aligned}$$

внизу слева

$$\mathbf{CM} - \mathbf{DD}^{-1}\mathbf{CM} = \mathbf{CM} - \mathbf{CM} = \mathbf{0}$$

внизу справа

$$-\mathbf{CMBD}^{-1} + \mathbf{DD}^{-1} + \mathbf{DD}^{-1}\mathbf{CMBD}^{-1} = \mathbf{DD}^{-1} = \mathbf{I}.$$

Таким образом, правая часть также равна единичной матрице.

3.19. Сначала возьмем совместное распределение $p(\mathbf{x}_a; \mathbf{x}_b; \mathbf{x}_c)$ и проводим его маргинализацию, чтобы получить распределение $p(\mathbf{x}_a; \mathbf{x}_b)$. Используя результаты раздела 3.2.5, вновь получаем гауссово распределение со средним значением и ковариацией, заданными формулами

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

Из раздела 3.2.4 следует, что распределение $p(\mathbf{x}_a; \mathbf{x}_b)$ является гауссовым со средним значением и ковариацией, заданными формулами (3.65) и (3.66) соответственно.

3.20. Умножение левой части (3.210) на $(\mathbf{A} + \mathbf{BCD})$ дает простую единичную матрицу $\mathbf{I}$. В правой части получаем

$$\begin{aligned}
&(\mathbf{A} + \mathbf{BCD})(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}) \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\qquad - \mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BC}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BCDA}^{-1} = \mathbf{I}.
\end{aligned}$$

3.21. Из $\mathbf{y} = \mathbf{x} + \mathbf{z}$ очевидно, что $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{z}]$. Для ковариации получаем

$$\begin{aligned}
\mathrm{cov}[\mathbf{y}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] \\
&= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}] + \mathbb{E}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}] \\
&\quad + \underbrace{\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}}_{=0} + \underbrace{\mathbb{E}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}]}_{=0} \\
&= \mathrm{cov}[\mathbf{x}] + \mathrm{cov}[\mathbf{z}],
\end{aligned}$$

где использована независимость $\mathbf{x}$ и $\mathbf{z}$, а также $\mathbb{E}[(\mathbf{x} - \mathbf{E}[\mathbf{x}])] = \mathbb{E}[(\mathbf{z} - \mathbb{E}[\mathbf{z}])] = 0$, чтобы установить третье и четвертое слагаемые в разложении равными нулю. В случае одномерных переменных ковариации становятся дисперсиями, и в итоге получается результат упражнения 2.10 как частный случай.

3.22. Для маргинального распределения $p(\mathbf{x})$ из (3.76) видно, что среднее значение задается верхней частью (3.92), которая просто равна $\boldsymbol{\mu}$. Точно так же из (3.77) видно, что ковариация задается верхней левой частью (3.89) и поэтому равна $\boldsymbol{\Lambda}^{-1}$.

Теперь рассмотрим условное распределение $p(\mathbf{y}|\mathbf{x})$. Применяя результат (3.65) для условного среднего значения, получаем

$$\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{A}\mathbf{x} + \mathbf{b}.$$

Точно так же, применяя результат (3.66) для ковариации условного распределения, получаем

$$\mathrm{cov}[\mathbf{y}|\mathbf{x}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}\mathbf{T} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{A}\mathbf{T} = \mathbf{L}^{-1},$$

как и требовалось.

3.23. Для начала определим

$$\mathbf{X} = \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} \tag{112}$$

и

$$\mathbf{W} = -\mathbf{L}\mathbf{A}, \text{ и, таким образом, } \mathbf{W}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}, \tag{113}$$

поскольку $\mathbf{L}$ является симметричной. Используем (112) и (113), чтобы переписать (3.88) в виде

$$\mathbf{R} = \begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix},$$

и с помощью (3.60) получаем

$$\begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \\ -\mathbf{L}^{-1}\mathbf{W}\mathbf{M} & \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \end{pmatrix},$$

где теперь

$$\mathbf{M} = (\mathbf{X} - \mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{W})^{-1}.$$

Заменяя $\mathbf{X}$ и $\mathbf{W}$ с помощью (112) и (113) соответственно, получаем

$$\mathbf{M} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1} = \boldsymbol{\Lambda}^{-1},$$

$$-\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}$$

и

$$\begin{aligned} \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} &= \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \\ &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}, \end{aligned}$$

как и требовалось.

3.24. Подставляя левое выражение из (3.89) вместо $\mathbf{R}^{-1}$ в (3.91), получаем

$$\begin{aligned}
&\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{S}\mathbf{b} \end{pmatrix} \\
&\quad = \begin{pmatrix} \boldsymbol{\Lambda}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}) + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}) + (\mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})\mathbf{S}\mathbf{b} \end{pmatrix} \\
&\quad = \begin{pmatrix} \boldsymbol{\mu} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \end{pmatrix} \\
&\quad = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{b} \end{pmatrix}.
\end{aligned}$$

3.25. Поскольку $\mathbf{y} = \mathbf{x} + \mathbf{z}$, условное распределение $\mathbf{y}$ при заданном $\mathbf{x}$ можно записать в виде $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathrm{z}} + \mathbf{x}, \boldsymbol{\Sigma}_{\mathrm{z}})$. Это позволяет разложить совместное распределение $\mathbf{x}$ и $\mathbf{y}$ в виде $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, где $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\Sigma}_{\mathrm{x}})$. Таким образом, это принимает форму (3.83) и (3.84), где можем идентифицировать $\boldsymbol{\mu} \to \boldsymbol{\mu}_{\mathrm{x}}$, $\boldsymbol{\Lambda}^{-1} \to \boldsymbol{\Sigma}_{\mathrm{x}}$, $\mathbf{A} \to \mathbf{I}$, $\mathbf{b} \to \boldsymbol{\mu}_{\mathrm{z}}$ и $\mathbf{L}^{-1} \to \boldsymbol{\Sigma}_{\mathrm{z}}$. Теперь мы можно получить маргинальное распределение $p(\mathbf{y})$ с помощью результата (3.99), из которого получаем $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathrm{x}} + \boldsymbol{\mu}_{\mathrm{z}}, \boldsymbol{\Sigma}_{\mathrm{z}} + \boldsymbol{\Sigma}_{\mathrm{x}})$. Таким образом, как средние значения, так и ковариации являются аддитивными, что согласуется с результатами упражнения 3.21.

3.26. Квадратичная форма в экспоненте совместного распределения задается формулой

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \quad (114)$$

Теперь извлечем все члены, содержащие $\mathbf{x}$, и соберем их в стандартную гауссову квадратичную форму, дополняя квадрат:

$$\begin{aligned}
&= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^{\mathrm{T}}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})] + \text{const} \\
&= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \\
&\quad + \frac{1}{2}\mathbf{m}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{m} + \text{const}, \qquad (115)
\end{aligned}$$

где

$$\mathbf{m} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})].$$

Теперь можно выполнить интегрирование по $\mathbf{x}$, что устраняет первый член в (115). Затем извлекаем члены в $\mathbf{y}$ из последнего члена в (115) и объединяем их с остальными членами из квадратичной формы (114), которые зависят от $\mathbf{y}$, чтобы получить

$$= -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\{\mathbf{L} - \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\}\mathbf{y}$$
$$+ \mathbf{y}^{\mathrm{T}}[\{\mathbf{L} - \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\}\,\mathbf{b}$$
$$+ \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}]. \tag{116}$$

Точность маргинального распределения $p(\mathbf{y})$ можно определить по члену второго порядка в $\mathbf{y}$. Чтобы найти соответствующую ковариацию, берем обратное значение точности и применяем формулу инверсии Вудбери (Woodbury inversion) (3.210), чтобы получить

$$\{\mathbf{L} - \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\}^{-1} = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}, \tag{117}$$

что соответствует (3.94).

Далее определим среднее значение $\boldsymbol{\nu}$ маргинального распределения. Для этого воспользуемся (117) в (116), а затем доведем квадрат до полного квадрата, чтобы получить

$$-\frac{1}{2}(\mathbf{y} - \boldsymbol{\nu})\mathrm{T}(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{AT})^{-1}(\mathbf{y} - \boldsymbol{\nu}) + \text{const},$$

где

$$\boldsymbol{\nu} = (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})[(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b} + \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}].$$

Теперь рассмотрим два слагаемых в квадратных скобках, первое из которых содержит $\mathbf{b}$, а второе содержит $\boldsymbol{\mu}$. Первое из этих слагаемых дает просто $\mathbf{b}$, а слагаемое в $\boldsymbol{\mu}$ можно записать как

$$= (\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})\mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}$$
$$= \mathbf{A}(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{LA})(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{LA})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu},$$

где используем общий результат $(\mathbf{BC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}$. Таким образом, получаем (3.93).

3.27. Чтобы найти условное распределение $p(\mathbf{x}|\mathbf{y})$, начнем с квадратичной формы (114), соответствующей совместному распределению $p(\mathbf{x}; \mathbf{y})$. Однако теперь будем рассматривать $\mathbf{y}$ как константу и просто доведем квадрат по $\mathbf{x}$ до полного квадрата, чтобы получить

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^{\mathrm{T}}\,\mathbf{L}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})$$
$$= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})\mathbf{x} + \mathbf{x}^{\mathrm{T}}\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{AL}(\mathbf{y} - \mathbf{b})\} + \text{const}$$
$$= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA})(\mathbf{x} - \mathbf{m}),$$

где, как и в решении упражнения 3.26, мы определяем

$$\mathbf{m} = (\mathbf{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\{\mathbf{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\},$$

из чего непосредственно получаем среднее значение и ковариацию условного распределения в виде (3.95) и (3.96).

3.28. Дифференцируя (3.102) по $\mathbf{\Sigma}$, получаем два слагаемых:

$$-\frac{N}{2}\frac{\partial}{\partial\mathbf{\Sigma}}\ln|\mathbf{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial\mathbf{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

Для первого слагаемого можно непосредственно применить (A.28) и получить

$$-\frac{N}{2}\frac{\partial}{\partial\mathbf{\Sigma}}\ln|\mathbf{\Sigma}| = -\frac{N}{2}(\mathbf{\Sigma}^{-1})^{\mathrm{T}} = -\frac{N}{2}\mathbf{\Sigma}^{-1}.$$

Для второго слагаемого сначала перепишем сумму

$$\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = N\mathrm{Tr}[\mathbf{\Sigma}^{-1}\mathbf{S}],$$

где

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}.$$

Используя это выражение вместе с (A.21), где $x = \Sigma_{ij}$ (элемент $(i; j)$ в $\mathbf{\Sigma}$), и свойствами следа, получаем

$$\begin{aligned}
\frac{\partial}{\partial\Sigma_{ij}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) &= N\frac{\partial}{\partial\Sigma_{ij}}\mathrm{Tr}[\mathbf{\Sigma}^{-1}\mathbf{S}] \\
&= N\mathrm{Tr}\left[\frac{\partial}{\partial\Sigma_{ij}}\mathbf{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\frac{\partial\mathbf{\Sigma}}{\partial\Sigma_{ij}}\mathbf{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\frac{\partial\mathbf{\Sigma}}{\partial\Sigma_{ij}}\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}\right] \\
&= -N(\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1})_{ij},
\end{aligned}$$

где использовано (A.26). Обратите внимание, что в последнем шаге проигнорирован тот факт, что $\Sigma_{ij} = \Sigma_{ji}$, так что $\partial\mathbf{\Sigma}/\partial\Sigma_{ij}$ дает 1 только в позиции $(i; j)$, а во всех остальных позициях дает 0.

Тем не менее, считая этот результат действительным, получаем

$$-\frac{1}{2}\frac{\partial}{\partial\mathbf{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \frac{N}{2}\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}.$$

Объединяя производные двух слагаемых и приравнивая результат к нулю, получаем

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Перегруппировка дает

$$\boldsymbol{\Sigma} = \mathbf{S},$$

как и требовалось.

3.29. Вывод (3.46) следует непосредственно из обсуждения, приведенного в тексте книги между (3.42) и (3.46). Если $m = n$, то, используя (3.46), получаем $\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$, тогда как если $n \neq m$, то две точки данных $\mathbf{x}_n$ и $\mathbf{x}_m$ являются независимыми, и, следовательно, $\mathbb{E}[\mathbf{x}_n\mathbf{x}_n] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$, где использовано (3.42). Объединяя эти результаты, получаем (3.213). Из (3.42) и (3.46) получаем

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\mathbf{x}_n^{\mathrm{T}} - \frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \frac{2}{N}\mathbf{x}_n\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}\right] \\
&= \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right\} \\
&= \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma}, \qquad (118)
\end{aligned}$$

как и требовалось.

3.30. Используя соотношение (3.214), получаем

$$1 = \exp(iA)\exp(-iA) = (\cos A + i\sin A)(\cos A - i\sin A) = \cos^2 A + \sin^2 A.$$

Точно так же получаем

$$\begin{aligned}
\cos(A-B) &= \Re\exp\{i(A-B)\} \\
&= \Re\exp(iA)\exp(-iB) \\
&= \Re(\cos A + i\sin A)(\cos B - i\sin B) \\
&= \cos A\cos B + \sin A\sin B.
\end{aligned}$$

Наконец,

$$\begin{aligned}
\sin(A-B) &= \Im\exp\{i(A-B)\} \\
&= \Im\exp(iA)\exp(-iB) \\
&= \Im(\cos A + i\sin A)(\cos B - i\sin B) \\
&= \sin A\cos B - \cos A\sin B.
\end{aligned}$$

3.31. В выражении через ξ распределение фон Мизеса принимает вид

$$p(\xi) \propto \exp\{m\cos(m^{-1/2}\xi)\}.$$

Для больших значений m имеем $\cos(m^{-1/2}\xi) = 1 - m^{-1}\xi^2/2 + O(m^{-2})$ и, следовательно,

$$p(\xi) \propto \exp\{-\xi^2/2\},$$

и, таким образом, $p(\theta) \propto \exp\{-m(\theta - \theta_0)^2/2\}$.

3.32. Используя (3.133), можно записать (3.132) в виде

$$\sum_{n=1}^{N}(\cos\theta_0\sin\theta_n - \cos\theta_n\sin\theta_0) = \cos\theta_0\sum_{n=1}^{N}\sin\theta_n - \sin\theta_0\sum_{n=1}^{N}\cos\theta_n = 0.$$

Перегруппировав это, получаем

$$\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} = \frac{\sin\theta_0}{\cos\theta_0} = \tan\theta_0,$$

и теперь это можно решить относительно θ_0, чтобы получить (3.134).

3.33. Дифференцируя распределение фон Мизеса (3.129), получаем

$$p'(\theta) = -\frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta - \theta_0)\}\sin(\theta - \theta_0),$$

что исчезает, когда $\theta = \theta_0$ или когда $\theta = \theta_0 + \pi(\mathrm{mod}\,2\pi)$. Дифференцируя снова, получаем

$$p''(\theta) = -\frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta - \theta_0)\}[\sin^2(\theta - \theta_0) + \cos(\theta - \theta_0)].$$

Поскольку $I_0(m) > 0$, очевидно, что $p''(\theta) < 0$ при $\theta = \theta_0$, что, следовательно, представляет собой максимум плотности, в то время как $p''(\theta) > 0$ при $\theta = \theta_0 + \pi(\mathrm{mod}\,2\pi)$, что, следовательно, является минимумом.

3.34. Из (3.119) и (3.134) видно, что $\theta = \theta_0^{\mathrm{ML}}$. Используя это вместе с (3.118) и (3.127), можно переписать (3.137) следующим образом:

$$\begin{aligned} A(m_{\mathrm{ML}}) &= \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{\mathrm{ML}} + \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{\mathrm{ML}} \\ &= \bar{r}\cos\bar{\theta}\cos\theta_0^{\mathrm{ML}} + \bar{r}\sin\bar{\theta}\sin\theta_0^{\mathrm{ML}} \\ &= \bar{r}(\cos^2\theta_0^{\mathrm{ML}} + \sin^2\theta_0^{\mathrm{ML}}) \\ &= \bar{r}. \end{aligned}$$

3.35. Начиная с (3.26), аргумент экспоненты можно переписать как

$$-\frac{1}{2}\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}] + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

Последний член не зависит от $\mathbf{x}$, но зависит от $\boldsymbol{\mu}$ и $\boldsymbol{\Sigma}$, поэтому должен быть включен в $g(\eta)$. Второй член уже является внутренним произведением и может быть оставлен без изменений. Чтобы обработать первый член, определим D^2-мерные векторы $\mathbf{z}$ и $\boldsymbol{\lambda}$, которые состоят из столбцов $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ и $\boldsymbol{\Sigma}^{-1}$ соответственно, сложенных друг на друга. Теперь можно записать многомерное гауссово распределение в виде (3.138), где

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\lambda} \end{bmatrix},$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix},$$

$$h(\mathbf{x}) = (2\pi)^{-D/2},$$

$$g(\boldsymbol{\eta}) = |\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right).$$

3.36. Взяв первую производную от (3.172), получаем, как и в тексте,

$$-\nabla \ln g(\eta) = g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Взяв вновь градиент, получаем

$$\begin{aligned} -\nabla\nabla \ln g(\eta) &= g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}\mathrm{d}\mathbf{x} \\ &\quad + \nabla g(\eta)\int h(\mathbf{x})\exp\{\eta^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \mathbb{E}[\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}] - \mathbb{E}[\mathbf{u}(\mathbf{x})]\mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\ &= \mathrm{cov}[\mathbf{u}(\mathbf{x})], \end{aligned}$$

где использован результат (3.172).

3.37. Значение плотности $p(\mathbf{x})$ в точке $\mathbf{x}_n$ задается $h_{j(n)}$, где обозначение $j(n)$ означает, что точка данных $\mathbf{x}_n$ попадает в область j. Таким образом, функция логарифмической вероятности принимает вид

$$\sum_{n=1}^{N} \ln p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln h_{j(n)}.$$

Теперь необходимо учесть ограничение, согласно которому $p(\mathbf{x})$ должно интегрироваться до единицы. Поскольку $p(\mathbf{x})$ имеет постоянное значение h_i

в области i, объем которой равен Δ_i, ограничение нормализации принимает вид $\sum_i h_i \Delta_i = 1$. Введя множитель Лагранжа λ, далее минимизируем функцию

$$\sum_{n=1}^{N} \ln h_{j(n)} + \lambda \left(\sum_i h_i \Delta_i - 1 \right)$$

по отношению к h_k, чтобы получить

$$0 = \frac{n_k}{h_k} + \lambda \Delta_k,$$

где n_k обозначает общее количество точек данных, попадающих в область k. Умножив обе стороны на h_k, просуммировав по k и воспользовавшись ограничением нормализации, получаем $\lambda = -N$. Устранив λ, получаем окончательный результат для решения максимального правдоподобия для h_k в виде

$$h_k = \frac{n_k}{N} \frac{1}{\Delta_k}.$$

Отметим, что для ячеек одинакового размера $\Delta_k = \Delta$ получается высота ячейки h_k, которая пропорциональна доле точек, попадающих в эту ячейку, как и ожидалось.

3.38. Из (3.180) получаем

$$p(\mathbf{x}) = \frac{K}{NV(\rho)},$$

где $V(\rho)$ – это объем D-мерной гиперсферы с радиусом ρ, где, в свою очередь, ρ – это расстояние от $\mathbf{x}$ до его K-го ближайшего соседа в наборе данных. Таким образом, в полярных координатах, если рассматривать достаточно большие значения радиальной координаты r, получаем

$$p(\mathbf{x}) \propto r^{-D}.$$

Если рассмотреть интеграл $p(\mathbf{x})$ и обратить внимание, что элемент объема $\mathrm{d}\mathbf{x}$ можно записать как $r^{D-1}\mathrm{d}r$, то получим

$$\int p(\mathbf{x}) \mathrm{d}\mathbf{x} \propto \int_0^\infty r^{-D} r^{D-1} \mathrm{d}r = \int_0^\infty r^{-1} \mathrm{d}r,$$

который расходится логарифмически.

Глава 4. Однослойные сети: регрессия

4.1. Подставляя (1.1) в (1.2) и затем дифференцируя по w_i, получаем

$$\sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j x_n^j - t_n \right) x_n^i = 0. \tag{119}$$

Перегруппировав слагаемые, получаем требуемый результат.

4.2. Для регуляризованной функции суммы квадратов ошибок, заданной формулой (1.4), соответствующие линейные уравнения вновь получаются путем дифференцирования и принимают ту же форму, что и (4.53), но с заменой A_{ij} на $\tilde{A}_{ij}$, заданную формулой

$$\tilde{A}_{ij} = A_{ij} + \lambda I_{ij}. \tag{120}$$

4.3. Используя (4.6), получаем

$$\begin{aligned} 2\sigma(2a) - 1 &= \frac{2}{1 + e^{-2a}} - 1 \\ &= \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}} \\ &= \frac{1 - e^{-2a}}{1 + e^{-2a}} \\ &= \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \\ &= \tanh(a). \end{aligned}$$

Если теперь принять $a_j = (x - \mu_j)/2s$, можно переписать (4.57) как

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}(2\sigma(2a_j) - 1 + 1) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh(a_j), \end{aligned}$$

где $u_j = w_j/2$ для $j = 1, \ldots, M$ и $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

4.4. Сначала запишем

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v} &= \boldsymbol{\Phi}\tilde{\mathbf{v}} \\ &= \boldsymbol{\varphi}_1\tilde{v}^{(1)} + \boldsymbol{\varphi}_2\tilde{v}^{(2)} + \ldots + \boldsymbol{\varphi}_M\tilde{v}^{(M)}, \end{aligned}$$

где $\boldsymbol{\varphi}_m$ – это m-й столбец $\boldsymbol{\Phi}$ и $\tilde{\mathbf{v}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v}$. Сравнивая это с решением по методу наименьших квадратов в (4.14), получаем, что

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

соответствует проекции $\mathbf{t}$ на пространство, построенное столбцами $\boldsymbol{\Phi}$. Чтобы убедиться, что это действительно ортогональная проекция, сначала заметим, что для любого столбца $\boldsymbol{\Phi}$, $\boldsymbol{\varphi}_j$,

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\varphi}_j = [\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}]_j = \boldsymbol{\varphi}_j$$

и, следовательно,

$$(\mathbf{y} - \mathbf{t})^{\mathrm{T}}\boldsymbol{\varphi}_j = (\boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} - \mathbf{t})^{\mathrm{T}}\boldsymbol{\varphi}_j = \mathbf{t}^{\mathrm{T}}\,\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\,\boldsymbol{\Phi}^{\mathrm{T}} - \mathbf{I})^{\mathrm{T}}\boldsymbol{\varphi}_j = 0,$$

и, таким образом, $(\mathbf{y} - \mathbf{t})$ ортогонально каждому столбцу $\boldsymbol{\Phi}$ и, значит, ортогонально $\mathcal{S}$.

4.5. Если определить $\mathbf{R} = \mathrm{diag}(r_1, \ldots, r_N)$ как диагональную матрицу с весовыми коэффициентами, то можно записать весовую сумму квадратов функции затрат в виде

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}\mathbf{R}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}).$$

Установив производную по $\mathbf{w}$ равной нулю и выполнив перегруппировку, получаем

$$\mathbf{w}^{\star} = (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{t},$$

что сводится к стандартному решению (4.14) для случая $\mathbf{R} = \mathbf{I}$.

Если сравнить (4.60) с (4.9)–(4.11), то видно, что r_n можно рассматривать как параметр точности (обратной дисперсии), характерный для точки данных $(x_n; t_n)$, который либо заменяет, либо масштабирует β.

В качестве альтернативы можно рассматривать r_n как эффективное число повторяющихся наблюдений точки данных $(x_n; t_n)$; это становится особенно очевидным, если рассмотреть (4.60) с r_n, принимающим положительные целые значения, хотя это справедливо для любого $r_n > 0$.

4.6. Взяв градиент (4.26) по $\mathbf{w}$ и приравняв его к нулю, получаем

$$0 = -\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n)\}\boldsymbol{\varphi}(\mathbf{x}_n) + \lambda\mathbf{w}, \tag{121}$$

что можно переписать как

$$0 = -\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \lambda\mathbf{w}. \tag{122}$$

Произведя перегруппировку для решения по $\mathbf{w}$, получаем

$$\mathbf{w} = (\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}, \tag{123}$$

как и требовалось.

4.7. Сначала запишем функцию логарифмической вероятности, которая задается следующим образом:

$$\ln L(\mathbf{W}, \boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n)).$$

Прежде всего приравниваем производную по $\mathbf{W}$ к нулю и получаем

$$0 = -\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n))\boldsymbol{\varphi}(\mathbf{x}_n)^{\mathrm{T}}.$$

Умножив на $\boldsymbol{\Sigma}$, введя матрицу плана $\boldsymbol{\Phi}$ и матрицу целевых данных $\mathbf{T}$, получаем

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{W} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{T}.$$

Решая уравнение для $\mathbf{W}$, получаем (4.14), как и требовалось.

Решение максимального правдоподобия для $\boldsymbol{\Sigma}$ несложно найти, обратившись к стандартному результату из главы **??**, дающему

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_n))^{\mathrm{T}} -$$

как и требовалось. Поскольку ищем совместное максимальное значение по отношению к $\mathbf{W}$ и $\boldsymbol{\Sigma}$, видно, что в этом выражении появляется $\mathbf{W}_{\mathrm{ML}}$, как и в стандартном результате для безусловного гауссова распределения.

4.8. Ожидаемые квадратичные потери для векторной целевой переменной определяются следующим образом:

$$\mathbb{E}[L] = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}, \mathbf{x})\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t}.$$

Нам необходимо выбрать $\mathbf{y}(\mathbf{x})$ таким образом, чтобы минимизировать $\mathbb{E}[L]$. Это можно сделать формально с помощью вариационного исчисления, что дает

$$\frac{\delta\mathbb{E}[L]}{\delta\mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t})p(\mathbf{t}, \mathbf{x})\mathrm{d}\mathbf{t} = 0.$$

Решая уравнение для $\mathbf{y}(\mathbf{x})$ и используя правила суммирования и умножения вероятностей, получаем

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t}p(\mathbf{t}, \mathbf{x})\mathrm{d}\mathbf{t}}{\int p(\mathbf{t}, \mathbf{x})\mathrm{d}\mathbf{t}} = \int \mathbf{t}p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t},$$

что является условным средним значением $\mathbf{t}$ при условии $\mathbf{x}$. Для случая скалярной целевой переменной получается

$$y(\mathbf{x}) = \int \mathbf{t}\, p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t},$$

что эквивалентно (4.37).

4.9. Начнем с разложения квадрата в (**??**) аналогично одномерному случаю в уравнении, предшествующем (4.39),

$$\begin{aligned}\|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2 \\ &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 + (\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}])^{\mathrm{T}}(\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}) \\ &\quad + (\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]) + \|\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2.\end{aligned}$$

Следуя рассмотрению одномерного случая, теперь подставляем это в (4.64) и проводим интегрирование по $\mathbf{t}$. Снова перекрестный член исчезает, и получаем

$$\mathbb{E}[L] = \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 p(\mathbf{x})\mathrm{d}\mathbf{x} + \int \mathrm{var}[\mathbf{t}|\mathbf{x}]p(\mathbf{x})\mathrm{d}\mathbf{x},$$

из чего прямо следует, что функция $\mathbf{y}(\mathbf{x})$, минимизирующая $\mathbb{E}[L]$, определяется $\mathbb{E}[\mathbf{t}|\mathbf{x}]$.

4.10. Это упражнение является повторением упражнения 4.9.

4.11. Чтобы доказать нормализацию распределения (4.66), рассмотрим интеграл

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{|x|^q}{2\sigma^2}\right)\mathrm{d}x = 2\int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right)\mathrm{d}x$$

и произведем замену переменной

$$u = \frac{x^q}{2\sigma^2}.$$

Используя определение (**??**) гамма-функции, получаем

$$I = 2\int_0^{\infty} \frac{2\sigma^2}{q}(2\sigma^2 u)^{(1-q)/q}\exp(-u)\mathrm{d}u = \frac{2(2\sigma^2)^{1/q}\Gamma(1/q)}{q},$$

из чего вытекает нормализация (4.66).

Для данного распределения шума условное распределение целевой переменной при заданной входной переменной имеет вид

$$p(t|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}\exp\left(-\frac{|t - y(\mathbf{x}, \mathbf{w})|^q}{2\sigma^2}\right).$$

Функция правдоподобия получается путем умножения членов этой формы на все пары $\{\mathbf{x}_n, t_n\}$. Взяв логарифм и отбросив аддитивные константы, мы приходим к желаемому результату.

4.12. Поскольку $y(\mathbf{x})$ можно выбирать независимо для каждого значения $\mathbf{x}$, минимум ожидаемых потерь L_q можно найти путем минимизации подынтегрального выражения, заданного формулой

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\mathrm{d}t \tag{124}$$

для каждого значения $\mathbf{x}$. Приравнивание производной (124) по $y(\mathbf{x})$ к нулю дает условие стационарности

$$\int q|y(\mathbf{x}) - t|^{q-1}\text{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})\mathrm{d}t$$
$$= q\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t - q\int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t = 0,$$

которое также можно получить непосредственно, приравнивая функциональную производную (4.40) по $y(\mathbf{x})$ к нулю. Отсюда следует, что $y(\mathbf{x})$ должна удовлетворять

$$\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t. \tag{125}$$

В случае $q = 1$ это сводится к

$$\int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})\mathrm{d}t, \tag{126}$$

и это означает, что $y(\mathbf{x})$ должно быть условной медианой t.

Для $q \to 0$ следует отметить, что, как функция t, величина $|y(\mathbf{x}) - t|^q$ близка к 1 везде, кроме некоторой небольшой области вокруг $t = y(\mathbf{x})$, где она падает до нуля. Таким образом, значение (124) будет близко к 1, поскольку плотность $p(t)$ нормализована, но слегка уменьшена из-за «выемки» вблизи $t = y(\mathbf{x})$. Наибольшее уменьшение в (124) получится в случае выбора местоположения выемки таким образом, чтобы оно совпадало с наибольшим значением $p(t)$, т. е. с (обусловленной) модой.

Глава 5. Однослойные сети: классификация

5.1. Рассмотрим компоненту t_k целевого вектора $\mathbf{t}$. По определению имеет место

$$p(t_k = 1|\mathbf{x}) = p(\mathcal{C}_k|\mathbf{x}), \tag{127}$$

$$p(t_k = 0|\mathbf{x}) = 1 - p(\mathcal{C}_k|\mathbf{x}). \tag{128}$$

Взяв математическое ожидание t_k, получаем

$$\mathbb{E}[t_k|\mathbf{x}] = 1 \times p(\mathcal{C}_k|\mathbf{x}) + 0 \times \{1 - p(\mathcal{C}_k|\mathbf{x})\} = p(\mathcal{C}_k|\mathbf{x}), \tag{129}$$

как и требовалось.

5.2. Предположим, что выпуклые оболочки $\{\mathbf{x}_n\}$ и $\{\mathbf{y}_m\}$ пересекаются. Тогда существует по крайней мере одна точка $\mathbf{z}$, для которой

$$\mathbf{z} = \sum_n \alpha_n \mathbf{x}_n = \sum_m \beta_m \mathbf{y}_m, \tag{130}$$

где $\beta_m \geq 0$ для всех m и $\sum_m \beta_m = 1$. Если бы $\{\mathbf{x}_n\}$ и $\{\mathbf{y}_m\}$ также были линейно разделимы, то получилось бы, что

$$\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_n \alpha_n \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 = \sum_n \alpha_n(\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) > 0, \tag{131}$$

поскольку $\hat{\mathbf{w}}\mathbf{x}_n + w_0 > 0$, а $\{\alpha_n\}$ все неотрицательны и их сумма равна 1. Однако по соответствующему аргументу

$$\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_m \beta_m \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0 = \sum_m \beta_m(\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) < 0, \tag{132}$$

что является противоречием, и, следовательно, $\{\mathbf{x}_n\}$ и $\{\mathbf{y}_m\}$ не могут быть линейно разделимы, если их выпуклые оболочки пересекаются.

Если же вместо этого предположить, что $\{\mathbf{x}_n\}$ и $\{\mathbf{y}_m\}$ линейно разделимы, и рассмотреть точку $\mathbf{z}$ в пересечении их выпуклых оболочек, то возникает то же противоречие. Таким образом, такая точка не может существовать, и пересечение выпуклых оболочек $\{\mathbf{x}_n\}$ и $\{\mathbf{y}_m\}$ должно быть пустым.

5.3. Для целей данного упражнения явно покажем вклад весов смещения в (5.14), получив

$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\}, \tag{133}$$

где $\mathbf{w}_0$ – это вектор-столбец весов смещения (верхняя строка $\tilde{\mathbf{W}}$ в транспонированном виде), а $\mathbf{1}$ – это вектор-столбец из N элементов.

Можно взять производную (133) по $\mathbf{w}_0$, получая

$$2N\mathbf{w}_0 + 2(\mathbf{X}\mathbf{W} - \mathbf{T})^{\mathrm{T}}\mathbf{1}. \tag{134}$$

Установив это равным нулю и решая для $\mathbf{w}_0$, получаем

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}}, \tag{135}$$

где

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^{\mathrm{T}}\mathbf{1} \quad \text{и} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{1}. \tag{136}$$

Если подставить (135) в (133), получаем

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\{(\mathbf{X}\mathbf{W} + \bar{\mathbf{T}} - \bar{\mathbf{X}}\mathbf{W} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \bar{\mathbf{T}} - \bar{\mathbf{X}}\mathbf{W} - \mathbf{T})\}, \tag{137}$$

где

$$\bar{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}} \quad \text{и} \quad \bar{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}. \tag{138}$$

Устанавливая производную по $\mathbf{W}$ равной нулю, получаем

$$\mathbf{W} = (\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{T}} = \hat{\mathbf{X}}^{\dagger}\hat{\mathbf{T}}, \tag{139}$$

где определено $\hat{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}}$ и $\hat{\mathbf{T}} = \mathbf{T} - \bar{\mathbf{T}}$.

Теперь рассмотрим прогноз для нового вектора входа $\mathbf{x}^{\star}$,

$$\begin{aligned}\mathbf{y}(\mathbf{x}^{\star}) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^{\star} + \mathbf{w}_0 \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^{\star} + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\ &= \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}(\hat{\mathbf{X}}^{\dagger})^{\mathrm{T}}(\mathbf{x}^{\star} - \bar{\mathbf{x}}).\end{aligned} \tag{140}$$

Если применить (5.97) к $\bar{\mathbf{t}}$, то получим

$$\mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{1} = -b. \tag{141}$$

Следовательно, применив (5.97) к (140), получаем

$$\begin{aligned}\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}^{\star}) &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} + \mathbf{a}^{\mathrm{T}}\hat{\mathbf{T}}(\hat{\mathbf{X}}^{\dagger})^{\mathrm{T}}(\mathbf{x}^{\star} - \bar{\mathbf{x}}) \\ &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = -b,\end{aligned}$$

поскольку $\mathbf{a}^{\mathrm{T}}\hat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \bar{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

5.4. Если рассматривать несколько одновременных ограничений, то (5.97) принимает вид

$$\mathbf{A}\mathbf{t}_n + \mathbf{b} = \mathbf{0}, \tag{142}$$

где $\mathbf{A}$ – это матрица, а $\mathbf{b}$ – это столбец-вектор, такой, что каждый ряд $\mathbf{A}$ и элемент $\mathbf{b}$ соответствуют одному линейному ограничению.

Если применить (142) к (140), то получим

$$\begin{aligned}\mathbf{A}\mathbf{y}(\mathbf{x}^{\star}) &= \mathbf{A}\bar{\mathbf{t}} - \mathbf{A}\hat{\mathbf{T}}^{\mathrm{T}}(\hat{\mathbf{X}}^{\dagger})^{\mathrm{T}}(\mathbf{x}^{\star} - \bar{\mathbf{x}}) \\ &= \mathbf{A}\bar{\mathbf{t}} = -\mathbf{b},\end{aligned}$$

поскольку $\mathbf{A}\hat{\mathbf{T}}^{\mathrm{T}} = \mathbf{A}(\mathbf{T} - \bar{\mathbf{T}})^{\mathrm{T}} = \mathbf{b}\mathbf{1}^{\mathrm{T}} - \mathbf{b}\mathbf{1}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$. Таким образом, $\mathbf{A}\mathbf{y}(\mathbf{x}^{\star}) + \mathbf{b} = \mathbf{0}$.

5.5. С помощью определений (5.30) и (5.31) получаем:

$$\begin{aligned}\text{Точность (Precision)} \times \text{Отзыв (Recall)} &= \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \times \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \\ &= \frac{N_{\mathrm{TP}}^2}{(N_{\mathrm{TP}} + N_{\mathrm{FP}})(N_{\mathrm{TP}} + N_{\mathrm{FN}})}.\end{aligned}$$

Точно так же, принимая общий знаменатель, получаем:

$$\begin{aligned}\text{Точность (Precision)} \times \text{Отзыв (Recall)} &= \frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}} + \frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}} \\ &= \frac{N_{\text{TP}}(N_{\text{TP}}+N_{\text{FN}}) + N_{\text{TP}}(N_{\text{TP}}+N_{\text{FP}})}{(N_{\text{TP}}+N_{\text{FP}})(N_{\text{TP}}+N_{\text{FN}})} \\ &= \frac{N_{\text{TP}}(2N_{\text{TP}}+N_{\text{FN}}+N_{\text{FP}})}{(N_{\text{TP}}+N_{\text{FP}})(N_{\text{TP}}+N_{\text{FN}})}.\end{aligned}$$

Подставляя эти два результата в (5.38), получаем

$$F = \frac{2N_{\text{TP}}}{2N_{\text{TP}}+N_{\text{FP}}+N_{\text{FN}}}, \tag{143}$$

как и требовалось.

5.6. Поскольку функция квадратного корня является монотонной для неотрицательных чисел, можно взять квадратный корень из отношения $a \leq b$, чтобы получить $a^{1/2} \leq b^{1/2}$. Затем умножим обе части на неотрицательное число $a^{1/2}$, чтобы получить $a \leq (ab)^{1/2}$.

Вероятность ошибочной классификации, согласно (**??**), равна

$$\begin{aligned}p(\text{ошибка}) &= \int_{\mathcal{R}_1} p(\mathbf{x}; \mathcal{C}_2)\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}; \mathcal{C}_1)\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned} \tag{144}$$

Поскольку для минимизации вероятности ошибочной классификации была выбрана область принятия решений, необходимо, чтобы $p(\mathcal{C}_2|\mathbf{x}) \leq p(\mathcal{C}_1|\mathbf{x})$ в области $\mathcal{R}_1$ и $p(\mathcal{C}_1|\mathbf{x}) \leq p(\mathcal{C}_2|\mathbf{x})$ в области $\mathcal{R}_2$. Теперь применим результат $a \leq b \Rightarrow a^{1/2} \leq b^{1/2}$, чтобы получить

$$\begin{aligned}p(\text{ошибка}) &\leq \int_{\mathcal{R}_1} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\quad + \int_{\mathcal{R}_2} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int \{p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\}^{1/2}\mathrm{d}\mathbf{x},\end{aligned} \tag{145}$$

поскольку оба интеграла имеют одинаковую подынтегральную функцию. Конечный интеграл берется по всей области $\mathbf{x}$.

5.7. Подставляя $L_{kj} = 1 - \delta_{kj}$ в (5.23) и используя тот факт, что апостериорные вероятности в сумме равны единице, определяем, что для каждого $\mathbf{x}$ следует выбрать класс j, для которого $1 - p(\mathcal{C}_j|\mathbf{x})$ является минимальным, что эквивалентно выбору j, для которого апостериорная вероятность $p(\mathcal{C}_j|\mathbf{x})$ является максимальной. Эта матрица потерь присваивает потерю, равную единице, если пример классифицирован неправильно, и потерю, равную нулю, если

он классифицирован правильно, и, следовательно, минимизация ожидаемой потери минимизирует частоту неправильной классификации.

5.8. Из (5.23) видно, что для общей матрицы потерь и произвольных априорных вероятностей классов ожидаемые потери минимизируются путем отнесения входа $\mathbf{x}$ к классу j, который минимизирует

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k), \tag{146}$$

и, таким образом, имеется прямое соотношение между априорными вероятностями $p(\mathcal{C}_k)$ и матрицей потерь L_{kj}.

5.9. Рассмотрим сумму по точкам данных в (5.100) как приближение конечной выборки к ожидаемому значению, как показано в (2.40). Принимая предел $N \to \infty$, можно использовать (2.39), чтобы записать ожидаемое значение в виде

$$\mathbb{E}[p(\mathcal{C}_k|\mathbf{x})] = \int p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} = \int p(\mathcal{C}_k, \mathbf{x}) \mathrm{d}\mathbf{x} = p(\mathcal{C}_k), \tag{147}$$

где использованы правила произведения и суммы вероятностей.

5.10. Вектор $\mathbf{x}$ принадлежит классу $\mathcal{C}_k$ с вероятностью $p(\mathcal{C}_k|\mathbf{x})$. Если примем решение отнести $\mathbf{x}$ к классу $\mathcal{C}_j$, то понесем ожидаемый убыток $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$, тогда как если примем решение об отклонении, то понесем убыток λ. Таким образом, если

$$j = \operatorname*{argmin}_l \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}), \tag{148}$$

то ожидаемый убыток будет минимизирован в случае, если предпринять следующие действия:

$$\text{выбрать} \begin{cases} \text{класс } j, & \text{если } \min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) < \lambda; \\ \text{отклонить} & \text{в противном случае.} \end{cases} \tag{149}$$

Для матрицы потерь $L_{kj} = 1 - I_{kj}$ получаем $\sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_l|\mathbf{x})$ и, следовательно, отклоняем, если наименьшее значение $1 - p(\mathcal{C}_l|\mathbf{x})$ меньше, чем λ, или, что эквивалентно, если наибольшее значение $p(\mathcal{C}_l|\mathbf{x})$ меньше, чем $1 - \lambda$. В стандартном критерии отклонения принимается отклонение, если наибольшая апостериорная вероятность меньше θ. Таким образом, эти два критерия отклонения эквивалентны при условии, что $\theta = 1 - \lambda$.

5.11. Из (5.42) получаем:

$$\begin{aligned} 1 - \sigma(a) &= 1 - \frac{1}{1 + e^{-a}} = \frac{1 + e^{-a} - 1}{1 + e^{-a}} \\ &= \frac{e^{-a}}{1 + e^{-a}} = \frac{1}{e^{a} + 1} = \sigma(-a). \end{aligned}$$

Обратная функция логистической сигмоиды может быть легко определена следующим образом:

$$\begin{aligned} y = \sigma(a) &= \frac{1}{1+e^{-a}} \\ \Rightarrow \frac{1}{y} &= 1 = e^{-a} \\ \Rightarrow \ln\left\{\frac{1-y}{y}\right\} &= -a \\ \Rightarrow \ln\left\{\frac{y}{1-y}\right\} &= a = \sigma^{-1}(y). \end{aligned}$$

5.12. Подставляя (5.47) в (5.41), получаем, что нормализующие константы аннулируются, и в результате остается

$$\begin{aligned} a &= \ln \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)p(\mathcal{C}_1)}{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)p(\mathcal{C}_2)} \\ &= -\frac{1}{2}(\mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} - \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_1 \\ &\qquad - \mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} + \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. \end{aligned}$$

Подставляя это в правую часть формулы (5.40), получаем (5.48), где $\mathbf{w}$ и w_0 заданы формулами (5.49) и (5.50) соответственно.

5.13. Функция правдоподобия задается формулой

$$p(\{\varphi_n, \mathbf{t}_n\}|\{\pi_k\}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\{p(\varphi_n|\mathcal{C}_k)\pi_k\}^{t_{nk}},$$

и, взяв логарифм, получаем

$$\ln p(\{\varphi_n, \mathbf{t}_n\}|\{\pi_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln p(\varphi_n|\mathcal{C}_k) + \ln \pi_k\}. \tag{150}$$

Чтобы максимизировать логарифм вероятности по отношению к π_k, необходимо сохранить ограничение $\sum_k \pi_k = 1$. Это можно сделать путем ввода множителя Лагранжа λ и максимизации

$$\ln p(\{\varphi_n, \mathbf{t}_n\}|\{\pi_k\}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right). \tag{151}$$

Установив производную по π_k равной нулю, получаем

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0. \tag{152}$$

После перегруппировки получается

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k. \tag{153}$$

Суммируя обе части по k, находим, что $\lambda = -N$, и, используя это для устранения λ, получаем (5.101).

5.14. Если подставить (5.102) в (150), а затем использовать определение многомерного гауссова распределения (3.26), получим

$$\ln p(\{\varphi_n, \mathbf{t}_n\}|\{\pi_k\}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln|\boldsymbol{\Sigma}| + (\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varphi} - \boldsymbol{\mu})\}, \tag{154}$$

где опущены члены, не зависящие от $\{\boldsymbol{\mu}_k\}$ и $\boldsymbol{\Sigma}$.

Установив производную правой части (154) по $\boldsymbol{\mu}_k$, полученную с помощью (A.19), равной нулю, получаем

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k) = 0. \tag{155}$$

Используя (153), можно перегруппировать это уравнение, чтобы получить (5.103).

Переписав правую часть (154) как

$$-\frac{1}{2}b\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\varphi} - \boldsymbol{\mu}_k)^{\mathrm{T}}]\}, \tag{156}$$

можно использовать (A.24) и (A.28) для вычисления производной по $\boldsymbol{\Sigma}^{-1}$. Установив это равным нулю, получаем

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k}^{\mathrm{T}} t_{nk}\{\boldsymbol{\Sigma} - [\boldsymbol{\Sigma} - (\boldsymbol{\varphi}_n - \boldsymbol{\mu}_n)(\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\} = 0. \tag{157}$$

Вновь используя (153), можно перегруппировать это, чтобы получить (5.104), где $\mathbf{S}_k$ задается (5.105).

Обратите внимание, что, как и в упражнении 3.28, совсем не требуется, чтобы $\boldsymbol{\Sigma}$ была симметричной, здесь просто констатируется факт, что решение автоматически является симметричным.

5.15. Предположим, что обучающий набор состоит из точек данных $\mathbf{x}_n$, каждая из которых помечена соответствующим классом $\mathcal{C}_k$. Это позволяет под-

бирать параметры $\{\mu_{ki}\}$ для каждого класса независимо. Из (5.64) функция логарифмической вероятности для класса $\mathcal{C}_k$ задается следующим образом:

$$\ln p(\mathcal{D}|\mathcal{C}_k) = \sum_{n=1}^{N}\sum_{i=1}^{D}\{x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\}. \tag{158}$$

Устанавливая производную по μ_{ki} равной нулю, получаем

$$0 = \sum_{n=1}^{N}\left\{\frac{x_{ni}}{\mu_{ki}} - \frac{(1-x_{ni})}{(1-\mu_{ki})}\right\}. \tag{159}$$

Перегруппировав для решения по μ_{ki}, в конечном итоге получаем

$$\mu_{ki} = \frac{1}{N}\sum_{n=1}^{N} x_{ni}, \tag{160}$$

что является наглядным результирующим выражением, по которому для каждого класса k и каждого компонента i значение μ_{ki} определяется средним значением соответствующих компонентов x_{ni} тех векторов данных, которые принадлежат классу $\mathcal{C}_k$. Поскольку x_{ni} являются бинарными, это просто доля точек данных, для которых соответствующее значение i равно единице.

5.16. Генеративная модель для $\boldsymbol{\varphi}$, соответствующая выбранной схеме кодирования, задается формулой

$$p(\boldsymbol{\varphi}|\mathcal{C}_k) = \prod_{m=1}^{M} p(\boldsymbol{\varphi}_m|\mathcal{C}_k), \tag{161}$$

где

$$p(\boldsymbol{\varphi}_m|\mathcal{C}_k) = \prod_{l=1}^{L}\mu_{kml}^{\varphi_{ml}}, \tag{162}$$

где, в свою очередь, $\{\mu_{kml}\}$ являются параметрами многочленных моделей для $\boldsymbol{\varphi}$.

После подстановки этого выражения в (5.46) получается, что

$$\begin{aligned} a_k &= \ln p(\boldsymbol{\varphi}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M}\ln p(\boldsymbol{\varphi}_m|\mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M}\sum_{l=1}^{L}\varphi_{ml}\ln\mu_{kml}, \end{aligned} \tag{163}$$

что является линейным по φ_{ml}.

5.17. Обозначим набор данных как $\mathcal{D} = \{\varphi_{nml}\}$, где $n = 1, \ldots, N$. Исходя из предположения наивного Байеса, можно подобрать каждый класс $\mathcal{C}_k$ отдельно

к обучающим данным. Для класса $\mathcal{C}_k$ функция логарифмического правдоподобия принимает вид

$$\ln p(\mathcal{D}|\mathcal{C}_k) = \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{l=1}^{L}\varphi_{nml}\ln\mu_{kml}. \tag{164}$$

Следует отметить, что параметр μ_{kml} представляет вероятность того, что для класса $\mathcal{C}_k$ компонент m будет иметь ненулевой элемент в позиции l. Чтобы найти решение с максимумом правдоподобия, необходимо учитывать ограничение, согласно которому сумма этих вероятностей должна равняться единице, отдельно для каждого значения m, так что

$$\sum_{l=1}^{L}\mu_{kml} = 1. \tag{165}$$

С этим можно справиться путем введения множителей Лагранжа, по одному для каждого компонента, а затем путем максимизации модифицированной функции правдоподобия, заданной формулой

$$\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{l=1}^{L}\varphi_{nml}\ln\mu_{kml} + \sum_{m=1}^{M}\varphi_m\left(\sum_{l=1}^{L}\mu_{kml} - 1\right). \tag{166}$$

Приравнивание производной по μ_{kml} к нулю приводит к

$$0 = \sum_{n=1}^{N}\left\{\frac{\varphi_{nml}}{\mu_{kml}}\right\} + \lambda_m. \tag{167}$$

Сделав перегруппировку для решения по μ_{kml}, получаем

$$\mu_{kml} = -\frac{1}{\lambda_m}\sum_{n=1}^{N}\varphi_{nml}. \tag{168}$$

Чтобы найти множители Лагранжа, подставляем этот результат в ограничение (165) и перегруппировываем, получая

$$\lambda_m = -\sum_{n=1}^{N}\sum_{l=1}^{L}\varphi_{nml}. \tag{169}$$

Теперь используем это для замены множителя Лагранжа в (168), чтобы получить окончательный результат для решения максимального правдоподобия для параметров в виде

$$\mu_{kml} = \frac{\sum_{n=1}^{N}\varphi_{nml}}{\sum_{n=1}^{N}\sum_{l=1}^{L}\varphi_{nml}}. \tag{170}$$

5.18. Дифференцируя (5.42), получаем

$$\begin{aligned}\frac{d\sigma}{da} &= \frac{e^{-a}}{(1+e^{-a})^2} \\ &= \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} \\ &= \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} \\ &= \sigma(a)(1-\sigma(a)).\end{aligned}$$

5.19. Начнем с вычисления производной (5.74) по y_n:

$$\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \tag{171}$$

$$= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)}$$

$$= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \tag{172}$$

$$= \frac{y_n - t_n}{y_n(1-y_n)}. \tag{173}$$

Из (5.72) видно, что

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n). \tag{174}$$

Наконец, получаем

$$\nabla a_n = \boldsymbol{\varphi}_n, \tag{175}$$

где ∇ обозначает градиент относительно $\mathbf{w}$. Объединяя (173), (174) и (175) с помощью правила цепочки, получаем

$$\begin{aligned}\nabla E &= \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n \\ &= \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\varphi}_n,\end{aligned}$$

как и требовалось.

5.20. Если набор данных линейно разделим, любая граница принятия решения, разделяющая два класса, будет иметь свойство

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n \begin{cases} \geq 0, & \text{если } t_n = 1, \\ < 0 & \text{в противном случае.} \end{cases} \tag{176}$$

Кроме того, из (5.74) видно, что отрицательное логарифмическое правдоподобие будет минимизировано (т. е. правдоподобие будет максимизировано), когда $y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n) = t_n$ для всех n. Это будет иметь место, когда сигмоидная функция насыщена, что происходит в том случае, когда ее аргумент, $\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}$, стремится к $\pm\infty$, т. е. когда величина $\mathbf{w}$ стремится к бесконечности.

5.21. Исходя из (5.76), имеем

$$\frac{\partial y_k}{\partial a_k} = \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}}\right)^2 = y_k(1 - y_k),$$

$$\frac{\partial y_k}{\partial a_j} = -\frac{e^{a_k}e^{a_j}}{\left(\sum_i e^{a_i}\right)^2} = -y_k y_j, \qquad j \neq k.$$

Комбинируя эти результаты, получаем (5.78).

5.22. Из (5.80) получаем

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}. \tag{177}$$

Если объединить это с (5.78) с помощью правила цепочки, получим

$$\begin{aligned}\frac{\partial E}{\partial a_{nj}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \\ &= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \\ &= y_{nj} - t_{nj},\end{aligned}$$

где использовано условие, что $\forall n : \sum_k t_{nk} = 1$.

Если объединить это с (175), снова используя правило цепочки, то получим (5.81).

5.23. Рассмотрим два случая, когда $a \geq 0$ и $a < 0$, по отдельности. В первом случае можно воспользоваться (3.25), чтобы переписать (5.86) как

$$\begin{aligned}\Phi(a) &= \int_{-\infty}^{0} \mathcal{N}(\theta|0,1)\mathrm{d}\theta + \int_0^a \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)\mathrm{d}\theta \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^{a/\sqrt{2}} \exp(-u^2)\sqrt{2}\mathrm{d}u \\ &= \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\},\end{aligned} \tag{178}$$

где в последней строке использовано (5.87).

Когда $a < 0$, симметрия гауссова распределения дает

$$\Phi(a) = 1 - \Phi(-a). \tag{179}$$

Объединяя это с (178), получаем

$$\begin{aligned}\Phi(a) &= 1 - \frac{1}{2}\left\{1 + \text{erf}\left(-\frac{a}{\sqrt{2}}\right)\right\} \\ &= \frac{1}{2}\left\{1 + \text{erf}\left(\frac{a}{\sqrt{2}}\right)\right\},\end{aligned}$$

где использован тот факт, что функция erf является антисимметричной, т. е. $\text{erf}(-a) = -\text{erf}(a)$.

5.24. Из (5.72) вытекает, что

$$\begin{aligned}\left.\frac{\mathrm{d}\sigma}{\mathrm{d}a}\right|_{a=0} &= \sigma(0)(1 - \sigma(0)) \\ &= \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}.\end{aligned} \tag{180}$$

Поскольку производная кумулятивной функции распределения является просто соответствующей функцией плотности, из (5.86) получаем

$$\begin{aligned}\left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} &= \lambda \mathcal{N}(0|0, 1) \\ &= \lambda \frac{1}{\sqrt{2\pi}}.\end{aligned}$$

Приравнивая это к (180), можно увидеть, что

$$\lambda = \frac{\sqrt{2\pi}}{4}, \text{ или, что эквивалентно, } \lambda^2 = \frac{\pi}{8}. \tag{181}$$

Сравнение логистической сигмоидной функции и масштабированной пробит-функции проиллюстрировано на рис. 5.12.

Глава 6. Глубокие нейронные сети

6.1. В правой части (6.51) произведем замену переменных $u = r^2$, чтобы получить

$$\frac{1}{2} S_D \int_0^\infty e^{-u} u^{D/2-1} \mathrm{d}u = \frac{1}{2} S_D \Gamma(D/2), \tag{182}$$

где использовано определение (**??**) гамма-функции. В левой части (6.51) можно использовать (2.126), чтобы получить $\pi^{D/2}$. Приравнивая эти выражения, получаем желаемый результат (6.53).

Объем сферы радиуса 1 в D-мерном пространстве определяется интегрированием:

$$V_D = S_D \int_0^1 r^{D-1} \mathrm{d}r = \frac{S_D}{D}. \tag{183}$$

Для $D = 2$ и $D = 3$ получаем следующие результаты:

$$S_2 = 2\pi,\ S_3 = 4\pi,\ V_2 = \pi a^2,\ V_3 = \frac{4}{3}\pi a^3. \tag{184}$$

6.2. Объем куба равен $(2a)^D$. Совместив это с (6.53) и (6.54), получаем (6.55). Используя формулу Стирлинга (6.56) в (6.55), для больших значений D соотношение становится следующим:

$$\frac{\text{объем сферы}}{\text{объем куба}} = \left(\frac{\pi e}{2D}\right)^{D/2} \frac{1}{D}, \tag{185}$$

при этом оно стремится к 0 при $D \to \infty$. Расстояние от центра куба до середины одной из сторон равно a, поскольку именно в этой точке он соприкасается со сферой. Точно таким же образом расстояние до одного из углов по теореме Пифагора равно $a\sqrt{D}$. Таким образом, соотношение равно $\sqrt{D}$.

6.3. Поскольку $p(\mathbf{x})$ является радиально симметричным, оно будет примерно постоянным по всей оболочке радиуса r и толщины ϵ. Эта оболочка имеет объем $S_D r^{D-1}\epsilon$ и, поскольку $\|\mathbf{x}\|^2 = r^2$, имеем

$$\int_{\text{shell}} p(\mathbf{x})\mathrm{d}\mathbf{x} \simeq p(r) S_D r^{D-1}\epsilon, \tag{186}$$

из чего получается (6.58). Стационарные точки $p(r)$ можно определить путем дифференцирования:

$$\frac{\mathrm{d}}{\mathrm{d}r} p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right] \exp\left(-\frac{r}{2\sigma^2}\right) = 0. \tag{187}$$

Решая уравнение для r и используя $D \gg 1$, получаем $\hat{r} \simeq \sqrt{D}\sigma$.

Далее необходимо отметить, что

$$\begin{aligned} p(\hat{r} + \epsilon) &\propto (\hat{r} + \epsilon)^{D-1} \exp\left[-\frac{(\hat{r} + \epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\hat{r} + \epsilon)^2}{2\sigma^2} + (D-1)\ln(\hat{r} + \epsilon)\right]. \end{aligned} \tag{188}$$

Теперь разложим $p(r)$ вокруг точки $\hat{r}$. Поскольку это стационарная точка $p(r)$, необходимо сохранить члены до второго порядка. Используя разложение $\ln(1+x) = x - x^2/2 + O(x^3)$ вместе с $D \gg 1$, получаем (6.59).

Наконец, из (6.57) получается, что плотность вероятности в начале координат задается формулой

$$p(\mathbf{x} = \mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{1/2}},$$

а плотность в точке $\|\mathbf{x}\| = \hat{r}$ задается из (6.57) в виде

$$p(\|\mathbf{x}\| = \hat{r}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{D}{2}\right),$$

где использовано $\hat{r} \simeq \sqrt{D}\sigma$. Таким образом, отношение плотностей задается как $\exp(D/2)$.

6.4. С учетом определения функции tanh имеем

$$\begin{aligned}
\tanh(a) &= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\
&= \frac{1 - e^{-2a}}{1 + e^{-2a}} \\
&= \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}} \\
&= \frac{2}{1 + e^{-2a}} - 1 \\
&= 2\sigma(2a) - 1,
\end{aligned} \tag{189}$$

где используется определение сигмоидной функции в (6.60). После перегруппировки получаем

$$\sigma(a) = \frac{1}{2}(\tanh(a/2) + 1). \tag{190}$$

В случае функции активации логистической сигмоидной функции аргумент функции активации выходного узла в (6.11) задается следующим образом:

$$\sum_{j=0}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right). \tag{191}$$

Подставляя сигмоидную функцию с помощью (190), получаем

$$\sum_{j=0}^{M} \tilde{w}_{kj}^{(2)} \tanh\left(\sum_{i=0}^{D} \tilde{w}_{ji}^{(1)} x_i \right), \tag{192}$$

где определено

$$\tilde{w}_{kj}^{(2)} = \frac{1}{2} \tilde{w}_{kj}^{(2)}, \qquad j = 1, \ldots, M, \tag{193}$$

$$\tilde{w}_{k0}^{(2)} = \frac{1}{2} \tilde{w}_{k0}^{(2)} + \frac{1}{2}, \tag{194}$$

$$\tilde{w}_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)}, \tag{195}$$

что вновь принимает форму (6.11).

6.5. Используя определение логистической сигмоиды, функцию активации swish можно записать как

$$\mathrm{sw}(x) = \frac{x}{1 + \exp(-\beta x)}. \tag{196}$$

Эта функция изображена на рис. 2 для значений $\beta = 0{,}1$, $\beta = 1{,}0$ и $\beta = 10$. В пределе $\beta \to \infty$ можно рассматривать поведение функции swish отдельно для положительных и отрицательных значений x. Для $x > 0$ функция $\exp(-\beta x) \to 0$ и, следовательно, $\mathrm{sw}(x) \to x$, а для $x < 0$ функция $\exp(-\beta x) \to \infty$ и, следуя из этого, $\mathrm{sw}(x) \to 0$. Таким образом, в этом пределе функция swish становится функцией ReLU.

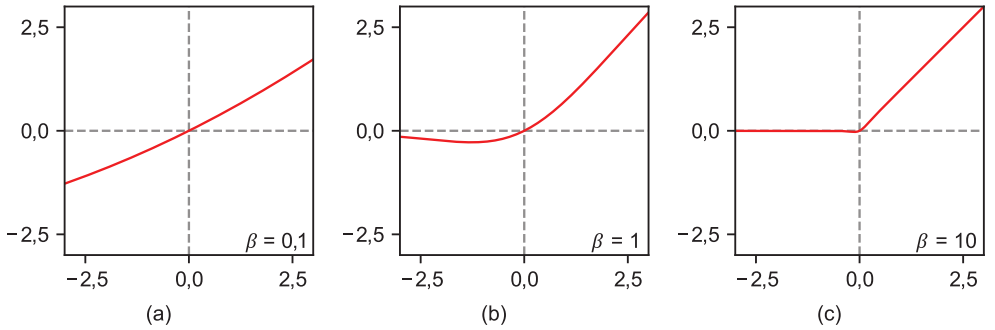

РИС. 2. *Функция активации swish, построенная для значений* $\beta = 0{,}1$, 1 *и* 10

6.6. Из (6.14), используя стандартные производные, получаем

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a}\tanh &= \frac{e^a}{e^a+e^{-a}} - \frac{e^a(e^a-e^{-a})}{(e^a+e^{-a})^2} + \frac{e^{-a}}{e^a+e^{-a}} + \frac{e^{-a}(e^a-e^{-a})}{(e^a+e^{-a})^2} \\ &= \frac{e^a+e^{-a}}{e^a+e^{-a}} + \frac{1-e^{2a}-e^{-2a}+1}{(e^a+e^{-a})^2} \\ &= 1 - \frac{e^{2a}-2+e^{-2a}}{(e^a+e^{-a})^2} \\ &= 1 - \frac{(e^a-e^{-a})(e^a-e^{-a})}{(e^a+e^{-a})(e^a+e^{-a})} \\ &= 1-\tanh^2(a).\end{aligned}$$

6.7. Функция активации softplus задается как

$$\zeta(a) = \ln(1+\exp(a)). \tag{197}$$

Можно доказать свойство (6.62), выполнив следующие шаги:

$$\zeta(a)-\zeta(-a) = \ln(1+\exp(a)) - \ln(1+\exp(-a)) \tag{198}$$

$$= \ln[\exp(a)(\exp(-a)+1)] - \ln(1+\exp(-a)) \tag{199}$$

$$= a + \ln(\exp(-a)+1) - \ln(1+\exp(-a)) \tag{200}$$

$$= a. \tag{201}$$

Чтобы доказать свойство (6.63), необходимо:

$$\ln\sigma(a) = -\ln(1+\exp(-a)) = -\zeta(-a). \tag{202}$$

Для производной (6.64) функции softplus имеем

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a}\zeta(a) &= \frac{\mathrm{d}}{\mathrm{d}a}\ln(1+\exp(a)) \\ &= \frac{\exp(a)}{1+\exp(a)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a).\end{aligned} \tag{203}$$

Наконец, чтобы найти обратную функцию (6.65) функции softplus, примем $y = \zeta^{-1}(a)$, тогда

$$a = \zeta(y) = \ln(1+\exp(y)). \tag{204}$$

После перегруппировки получаем

$$\exp(a) = 1+\exp(y) \tag{205}$$

и, следовательно,

$$y = \ln(\exp(a) - 1). \tag{206}$$

6.8. Дифференцируя ошибку (6.25) по σ^2 и приравнивая производную к нулю, получаем

$$0 = -\frac{1}{2\sigma^4}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\frac{1}{\sigma^2}. \tag{207}$$

Выполнив перегруппировку для решения по σ^2, получаем

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}^\star) - t_n\}^2, \tag{208}$$

как и требовалось.

6.9. Функция правдоподобия для независимых и идентично распределенных данных, $\{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_N, \mathbf{t}_N)\}$, при условном распределении (6.28) задается формулой

$$\prod_{n=1}^{N}\mathcal{N}(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}).$$

Если взять логарифм этого выражения, используя (3.26), то получим

$$\begin{aligned}
&\sum_{n=1}^{N}\ln\mathcal{N}(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}) \\
&\quad = -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^{\mathrm{T}}(\beta\mathbf{I})(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})) + \text{const} \\
&\quad = -\frac{\beta}{2}\sum_{n=1}^{N}||\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})||^2 + \text{const},
\end{aligned}$$

где const включает в себя члены, не зависящие от $\mathbf{w}$. Первый член в правой части пропорционален отрицательному значению (6.29), и, следовательно, максимизация логарифмического правдоподобия эквивалентна минимизации суммы квадратов ошибок.

6.10. В этом случае функция правдоподобия принимает вид

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma})\prod_{n=1}^{N}\mathcal{N}(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \boldsymbol{\Sigma})$$

с соответствующей функцией логарифмического правдоподобия

$$\ln p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) = -\frac{N}{2}(\ln|\boldsymbol{\Sigma}| + K\ln(2\pi)) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n), \tag{209}$$

где $\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n; \mathbf{w})$ и K – это размерность $\mathbf{y}$ и $\mathbf{t}$.

Если сначала рассматривать $\boldsymbol{\Sigma}$ как фиксированную и определенную, то из (209) можно исключить члены, не зависящие от $\mathbf{w}$, и, изменив знак, получить функцию ошибки

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n).$$

Если рассматривать максимизацию (209) по отношению к $\boldsymbol{\Sigma}$, то необходимо сохранить следующие слагаемые:

$$-\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n).$$

Переписав второй член, получаем

$$-\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\right].$$

Используя результаты из приложения **??**, можно максимизировать данное выражение, приравнивая производную по $\boldsymbol{\Sigma}^{-1}$ к нулю, что дает

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}.$$

Таким образом, оптимальное значение $\boldsymbol{\Sigma}$ зависит от $\mathbf{w}$ через $\mathbf{y}_n$.

Возможный способ решения этой взаимной зависимости между $\mathbf{w}$ и $\boldsymbol{\Sigma}$ при оптимизации – это использование итерационной схемы, чередующей обновления $\mathbf{w}$ и $\boldsymbol{\Sigma}$ до достижения некоторого критерия сходимости.

6.11. Пусть $t \in \{0, 1\}$ описывает метку набора данных, а $k \in \{0, 1\}$ определяет метку истинного класса. Необходимо, чтобы выход сети имел интерпретацию $y(\mathbf{x}, \mathbf{w}) = p(k = 1|\mathbf{x})$. Из правил вероятности следует, что

$$p(t = 1|\mathbf{x}) = \sum_{k=0}^{1} p(t = 1|k)p(k|\mathbf{x}) = (1 - \epsilon)y(\mathbf{x}, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}, \mathbf{w})).$$

Условная вероятность метки данных в таком случае равна

$$p(t|\mathbf{x}) = p(t = 1|\mathbf{x})^t(1 - p(t = 1|\mathbf{x})^{1-t}.$$

Формируя вероятность и взяв отрицательный логарифм, получаем функцию ошибки в виде

$$\begin{aligned}E(\mathbf{w}) = -\sum_{n=1}^{N}\{&t_n \ln[(1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))] \\ &+ (1 - t_n)\ln[1 - (1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) - \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))]\}.\end{aligned}$$

Смотрите также решение **??**.

6.12. Это просто соответствует масштабированию и сдвигу бинарных выходов, что непосредственно дает функцию активации с использованием обозначений из (**??**) в виде

$$y = 2\sigma(a) - 1.$$

Соответствующая функция ошибки может быть построена из (6.33) путем применения обратного преобразования к y_n и t_n, что дает

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N} \frac{1+t_n}{2} \ln \frac{1+y_n}{2} + \left(1 - \frac{1+t_n}{2}\right) \ln\left(1 - \frac{1+y_n}{2}\right) \\ &= -\frac{1}{2} \sum_{n=1}^{N} \{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n) + N \ln 2, \end{aligned}$$

где последний член можно опустить, поскольку он не зависит от $\mathbf{w}$.

Чтобы найти соответствующую функцию активации, достаточно просто применить линейное преобразование к логистической сигмоидальной функции, заданной (**??**), что дает

$$\begin{aligned} y(a) &= 2\sigma(a) - 1 = \frac{2}{1+e^{-a}} - 1 \\ &= \frac{1-e^{-a}}{1+e^{-a}} = \frac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} \\ &= \tanh(a/2). \end{aligned}$$

6.13. Для данной интерпретации $y_k(\mathbf{x}; \mathbf{w})$ условное распределение целевого вектора для нейронной сети с несколькими классами имеет вид

$$p(\mathbf{t} | \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{k=1}^{K} y_k^{t_k}.$$

Таким образом, для набора данных из N точек функция правдоподобия будет иметь вид

$$p(\mathbf{T} | \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}.$$

Взяв отрицательный логарифм для получения функции ошибки, получаем (6.36), как и требовалось. Здесь следует отметить, что это тот же результат, что и для логистической регрессионной модели с несколькими классами, которая определяется формулой (5.80).

6.14. Дифференцируя (6.33) по активации a_n, соответствующей конкретной точке данных n, получаем

$$\frac{\partial E}{\partial a_n} = -t_n \frac{1}{y_n} \frac{\partial y_n}{\partial a_n} + (1-t_n) \frac{1}{1-y_n} \frac{\partial y_n}{\partial a_n}. \tag{210}$$

Из (5.72) получаем

$$\frac{\partial y_n}{\partial a_n} = y_n(1 - y_n). \tag{211}$$

Подставляя (211) в (210), получаем

$$\begin{aligned}\frac{\partial E}{\partial a_n} &= -t_n \frac{y_n(1 - y_n)}{y_n} + (1 - t_n)\frac{y_n(1 - y_n)}{y_n} \\ &= y_n - t_n,\end{aligned}$$

как и требовалось.

6.15. Рассмотрим конкретную точку данных n и, чтобы минимизировать громоздкость, опустим суффикс n в таких переменных, как a_k и y_k. Используя правило цепной производной, можно записать

$$\frac{\partial E}{\partial a_k} = \sum_{j=1}^{K} \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial a_k}. \tag{212}$$

Из (6.36) получаем

$$\frac{\partial E}{\partial y_j} = -\frac{t_j}{y_j}. \tag{213}$$

Выражение (6.37) можно записать в виде

$$y_j = \frac{\exp(a_j)}{\sum_l \exp(a_l)}. \tag{214}$$

Для производной $\partial y_j/\partial a_k$ есть два компонента, один из числителя, а другой из знаменателя, так что

$$\begin{aligned}\frac{\partial y_j}{\partial a_k} &= \frac{\exp(a_j)\delta_{jk}}{\sum_l \exp(a_l)} - \frac{\exp(a_j)\exp(a_k)}{\left\{\sum_l \exp(a_l)\right\}^2} \\ &= y_j\delta_{jk} - y_j y_k.\end{aligned} \tag{215}$$

Подставляя (213) и (215) в (212), получаем

$$\frac{\partial E}{\partial a_k} = -\sum_{j=1}^{K} \frac{t_j}{y_j}\{y_j\delta_{jk} - y_j y_k\} = y_k - t_k, \tag{216}$$

как и требовалось. На последнем этапе использовалось

$$\sum_{j=1}^{K} t_j = 1, \tag{217}$$

что следует из схемы кодирования «1-из-K», используемой для $\{t_j\}$.

6.16. Исходя из стандартных тригонометрических правил, получаем положение конца первого плеча:

$$(x_1^{(1)}, x_2^{(1)}) = (L_1\cos(\theta_1), L_1\sin(\theta_1)).$$

Точно так же положение конца второго плеча относительно конца первого плеча задается соответствующим уравнением с угловым смещением π (см. рис. 6.16), которое равно изменению знака:

$$\begin{aligned}(x_1^{(2)}, x_2^{(2)}) &= (L_2\cos(\theta_1 + \theta_2 - \pi), L_1\sin(\theta_1 + \theta_2 - \pi)) \\ &= -(L_2\cos(\theta_1 + \theta_2), L_2\sin(\theta_1 + \theta_2)).\end{aligned}$$

Сложив все это вместе, следует также учесть, что θ_2 измеряется относительно первого плеча, и, таким образом, получается положение конца второго плеча относительно точки крепления первого плеча в виде

$$(x_1, x_2) = (L_1\cos(\theta_1) - L_2\cos(\theta_1 + \theta_2), L_1\sin(\theta_1) - L_2\sin(\theta_1 + \theta_2)).$$

6.17. Интерпретация γ_{nk} как апостериорной вероятности следует из теоремы Байеса для вероятности компонента, индексированного k, при заданных наблюдаемых данных $\mathbf{t}$, в которых все величины также обусловлены входной переменной $\mathbf{x}$. Поэтому $\mathbf{x}$ просто появляется как обусловливающая переменная в правой части всех величин. Из теоремы Байеса получаем

$$p(k|\mathbf{t}, \mathbf{x}) = \frac{p(\mathbf{t}|k, \mathbf{x})p(k|\mathbf{x})}{p(\mathbf{t}|\mathbf{x})}, \tag{218}$$

где, как обычно, знаменатель может быть выражен в виде маргинализации по членам в числителе, так что

$$p(\mathbf{t}|\mathbf{x}) = \sum_l p(\mathbf{t}|l, \mathbf{x})p(l|\mathbf{x}). \tag{219}$$

Величины $\pi_k(\mathbf{x})$, определенные формулой (6.40), удовлетворяют формуле (6.39) и, следовательно, отвечают требуемым условиям для того, чтобы рассматриваться как вероятности, поэтому можно приравнять $p(k|\mathbf{x}) = \pi_k(\mathbf{x})$. Аналогично условное распределение по классам $p(\mathbf{t}|k, \mathbf{x})$ задается гауссовой функцией $\mathcal{N}_{nk} = \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n), \sigma_k^2(\mathbf{x}_n))$. Подставляя в (218), получаем

$$p(k|\mathbf{t}_n, \mathbf{x}_n) = \frac{\pi_k \mathcal{N}_{nk}}{\sum_l \pi_l \mathcal{N}_{nl}} = \gamma_{nk}, \tag{220}$$

как и требовалось.

6.18. Начнем с использования правила цепочки, чтобы записать

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{221}$$

Следует отметить, что из-за связи между выходами, вызванной функцией активации softmax, зависимость от активации одного выходного блока затрагивает все выходные блоки.

Для первого фактора внутри суммы в правой части (221) стандартные производные, примененные к n-му члену (6.43), дают

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{222}$$

Для второго фактора из (5.78) получаем

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j (I_{jk} - \pi_k). \tag{223}$$

Объединяя (221), (222) и (223), получаем

$$\begin{aligned} \frac{\partial E_n}{\partial a_k^\pi} &= -\sum_{j=1}^{K} \frac{\gamma_{nj}}{\pi_j} \pi_j (I_{jk} - \pi_k) \\ &= -\sum_{j=1}^{K} \gamma_{nj} (I_{jk} - \pi_k) = -\gamma_{nj} + \sum_{j=1}^{K} \gamma_{nj} \pi_k = \pi_k - \gamma_{nk}, \end{aligned}$$

где используем тот факт, что, согласно (6.44), $\sum_{j=1}^{K} \gamma_{nj} = 1$ для всех n.

6.19. Примечание (авторов): см. решение 6.18.

Из (6.42) имеем

$$a_{kl}^\mu = \mu_{kl},$$

и, следовательно,

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \frac{\partial E_n}{\partial \mu_{kl}}.$$

Из (3.26), (6.43) и (6.44) получаем

$$\begin{aligned} \frac{\partial E_n}{\partial \mu_{kl}} &= -\frac{\pi_k \mathcal{N}_{nk}}{\sum_{k'} \pi_{k'} \mathcal{N}_{nk'}} \frac{t_{nl} - \mu_{kl}}{\sigma_k^2(\mathbf{x}_n)} \\ &= \gamma_{nk}(\mathbf{t}_n | \mathbf{x}_n) \frac{\mu_{kl} - t_{nl}}{\sigma_k^2(\mathbf{x}_n)}. \end{aligned}$$

6.20. Из (6.41) и (6.43) становится понятно, что

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k}\frac{\partial \sigma_k}{\partial a_k^\sigma}, \tag{224}$$

где из (6.41)

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \sigma_k. \tag{225}$$

Из (3.26), (6.43) и (6.44) получаем

$$\begin{aligned}\frac{\partial E_n}{\partial \sigma_k} &= -\frac{1}{\sum_{k'} \mathcal{N}_{nk'}}\left(\frac{L}{2\pi}\right)^{L/2}\left\{-\frac{L}{\sigma^{L+1}}\exp\left(-\frac{||\mathbf{t}_n - \boldsymbol{\mu}_k||^2}{2\sigma_k^2}\right)\right. \\ &\quad \left. + \frac{1}{\sigma^L}\exp\left(-\frac{||\mathbf{t}_n - \boldsymbol{\mu}_k||^2}{2\sigma_k^2}\right)\frac{||\mathbf{t}_n - \boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\} \\ &= \gamma_{nk}\left(\frac{L}{\sigma_k} - \frac{||\mathbf{t}_n - \boldsymbol{\mu}_k||^2}{\sigma_k^3}\right).\end{aligned}$$

Объединяя это с (224) и (225), получаем

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk}\left(L - \frac{||\mathbf{t}_n - \boldsymbol{\mu}_k||^2}{\sigma_k^2}\right).$$

6.21. Из (3.42) и (6.38) имеем

$$\begin{aligned}\mathbb{E}[\mathbf{t}|\mathbf{x}] &= \int \mathbf{t} p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t} \\ &= \int \mathbf{t}\sum_{k=1}^K \pi_k(\mathbf{x})\mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x}))\mathrm{d}\mathbf{t} \\ &= \sum_{k=1}^K \pi_k(\mathbf{x})\int \mathbf{t}\mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x}))\mathrm{d}\mathbf{t} \\ &= \sum_{k=1}^K \pi_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x}).\end{aligned}$$

Теперь введем сокращенную запись

$$\bar{\mathbf{t}}_k = \boldsymbol{\mu}_k(\mathbf{x}) \quad \text{и} \quad \bar{\mathbf{t}} = \sum_{k=1}^K \pi_k(\mathbf{x})\bar{\mathbf{t}}_k.$$

Используя это вместе с (3.42), (3.46), (6.38) и (6.48), получаем

$$\begin{aligned} s^2(\mathbf{x}) &= \mathbb{E}[\|\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2|\mathbf{x}] = \int \|\mathbf{t} - \overline{\mathbf{t}}\|^2 p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t} \\ &= \int (\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\overline{\mathbf{t}} - \overline{\mathbf{t}}^{\mathrm{T}}\mathbf{t} + \overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}) \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x}))\mathrm{d}\mathbf{t} \\ &= \sum_{k=1}^{K} \pi_k(\mathbf{x})\{\sigma_k^2 + \overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}}_k - \overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}} - \overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}_k + \overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}\} \\ &= \sum_{k=1}^{K} \pi_k(\mathbf{x})\{\sigma_k^2 + \|\overline{\mathbf{t}}_k - \overline{\mathbf{t}}\|^2\} \\ &= \sum_{k=1}^{K} \pi_k(\mathbf{x})\left\{\sigma_k^2 + \left\|\boldsymbol{\mu}_k(\mathbf{x}) - \sum_{l}^{K} \pi_l \boldsymbol{\mu}_l(\mathbf{x})\right\|^2\right\}. \end{aligned}$$

Глава 7. Градиентный спуск

7.1. Подставляя (7.10) в (7.7), получаем

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\sum_i \alpha_i \mathbf{u}_i^{\mathrm{T}}\right)\mathbf{H}\left(\sum_j \alpha_j \mathbf{u}_j\right).$$

Используя (7.8), получаем

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\sum_i \alpha_i \mathbf{u}_i^{\mathrm{T}}\right)\left(\sum_j \lambda_j \alpha_j \mathbf{u}_j\right).$$

Теперь, используя (7.9), получаем

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \alpha_i \sum_j \lambda_j \alpha_j \delta_{ij} \\ &= E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2, \end{aligned} \tag{226}$$

как и требовалось.

7.2. Из (7.8) и (7.10) получаем

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{H}\mathbf{u}_i = \mathbf{u}_i^{\mathrm{T}}\lambda_i\mathbf{u}_i = \lambda_i.$$

Предположим, что $\mathbf{H}$ является положительно определенной матрицей, так что (7.12) выполняется. Тогда, предположив $\mathbf{v} = \mathbf{u}_i$, получаем, что

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{H}\mathbf{u}_i > 0$$

для всех значений i. Таким образом, если $\mathbf{H}$ является положительно определенной матрицей, все ее собственные значения будут положительными.

И наоборот, предположим, что (**??**) выполняется. Тогда для любого вектора $\mathbf{v}$ можно использовать (7.13), чтобы получить

$$\begin{aligned}\mathbf{v}^{\mathrm{T}}\mathbf{H}\mathbf{v} &= \left(\sum_i c_i \mathbf{u}_i\right)^{\mathrm{T}} \mathbf{H}\left(\sum_j c_j \mathbf{u}_j\right) \\ &= \left(\sum_i c_i \mathbf{u}_i\right)^{\mathrm{T}} \left(\sum_j \lambda_j c_j \mathbf{u}_j\right) \\ &= \sum_i \lambda_i c_i^2 > 0,\end{aligned}$$

где были использованы (7.8) и (7.9) вместе с (**??**). Таким образом, если все собственные значения положительны, матрица Гессе будет положительно определенной.

7.3. Из (7.12) следует, что если $\mathbf{H}$ является положительно определенной матрицей, то второй член в (7.7) будет положительным всякий раз, когда $(\mathbf{w} - \mathbf{w}^\star)$ не равно нулю. Таким образом, наименьшее значение, которое может принимать $E(\mathbf{w})$, равно $E(\mathbf{w}^\star)$, и поэтому $\mathbf{w}^\star$ является минимумом $E(\mathbf{w})$. И наоборот, если $\mathbf{w}^\star$ является минимумом $E(\mathbf{w})$, то для любого вектора $\mathbf{w} \neq \mathbf{w}^\star$ будет справедливо $E(\mathbf{w}) > E(\mathbf{w}^\star)$. Это будет иметь место только в том случае, если второй член (7.7) будет положительным для всех значений $\mathbf{w} \neq \mathbf{w}^\star$ (поскольку первый член не зависит от $\mathbf{w}$). Поскольку $\mathbf{w} - \mathbf{w}^\star$ может быть задано любым вектором действительных чисел, из определения (7.12) следует, что $\mathbf{H}$ является положительно определенной.

7.4. Первые производные функции ошибки задаются следующим образом:

$$\frac{\partial E}{\partial w} = \sum_n (y_n - t_n) x_n, \tag{227}$$

$$\frac{\partial E}{\partial b} = \sum_n (y_n - t_n), \tag{228}$$

где $y_n = y(x_n, w, b)$. Вторые производные задаются следующим образом:

$$\frac{\partial^2 E}{\partial w^2} = \sum_n x_n^2, \tag{229}$$

$$\frac{\partial^2 E}{\partial w \partial b} = \frac{\partial^2 E}{\partial b \partial w} = \sum_n x_n = N\bar{x}, \tag{230}$$

$$\frac{\partial^2 E}{\partial b^2} = \sum_n 1 = N, \tag{231}$$

где $\bar{x}$ – это среднее значение выборки, определяемое как

$$\bar{x} = \frac{1}{N}\sum_n x_n.$$

Матрица Гессе задается следующим образом:

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial^2 E}{\partial w^2} & \dfrac{\partial^2 E}{\partial b \partial w} \\ \dfrac{\partial^2 E}{\partial w \partial b} & \dfrac{\partial^2 E}{\partial b^2} \end{pmatrix} = \begin{pmatrix} \sum_n x_n^2 & N\overline{x} \\ N\overline{x} & N \end{pmatrix}.$$

Следует отметить, что для этой простой модели матрица Гессе не зависит от целевых значений и от параметров модели. Она является функцией только от входных переменных данных. Трасса и определитель матрицы Гессе задаются следующим образом:

$$\mathrm{Tr}\,\mathbf{H} = \sum_n x_n^2 + N, \tag{232}$$

$$\det \mathbf{H} = N\sum_n x_n^2 - (N\overline{x})^2 = N^2\sigma^2, \tag{233}$$

где σ^2 – это дисперсия выборки, определяемая как

$$\sigma^2 = \frac{1}{N}\sum_n (x_n - \overline{x})^2.$$

Здесь видно, что след и определитель являются положительными. Поскольку определитель является произведением двух собственных значений матрицы Гессе, из этого следует, что либо оба собственных значения положительны, либо оба отрицательны. Поскольку след является суммой собственных значений, а след также положителен, из этого следует, что оба собственных значения должны быть положительными, а значит, матрица Гессе должна быть положительно определенной. Здесь не учитывается вырожденный случай, когда количество точек данных равно единице либо все точки данных одинаковы.

7.5. Отметим, что в оригинальном издании книги в правой части (7.64) пропущен знак минус (прим. авторов). Сначала следует обратить внимание на тот факт, что производная логистической сигмоидной функции задается формулой (5.18). Используя этот результат, первые производные функции ошибки задаются формулой:

$$\begin{aligned} \frac{\partial E}{\partial w} &= -\sum_n \left\{ \frac{t_n}{y_n} - \frac{1-t_n}{1-y_n} \right\} y_n(1-y_n)x_n \\ &= \sum_n (y_n - t_n)x_n, \end{aligned} \tag{234}$$

$$\begin{aligned} \frac{\partial E}{\partial b} &= -\sum_n \left\{ \frac{t_n}{y_n} - \frac{1-t_n}{1-y_n} \right\} y_n(1-y_n) \\ &= \sum_n (y_n - t_n), \end{aligned} \tag{235}$$

где $y_n = y(x_n, w, b)$. Вторые производные в этом случае задаются формулой:

$$\frac{\partial^2 E}{\partial w^2} = \sum_n y_n(1 - y_n)x_n^2, \tag{236}$$

$$\frac{\partial^2 E}{\partial w \partial b} = \frac{\partial^2 E}{\partial b \partial w} = \sum_n y_n(1 - y_n)x_n, \tag{237}$$

$$\frac{\partial^2 E}{\partial b^2} = \sum_n y_n(1 - y_n). \tag{238}$$

Матрица Гессе задается следующим образом:

$$\begin{aligned}\mathbf{H} &= \begin{pmatrix} \dfrac{\partial^2 E}{\partial w^2} & \dfrac{\partial^2 E}{\partial b \partial w} \\ \dfrac{\partial^2 E}{\partial w \partial b} & \dfrac{\partial^2 E}{\partial b^2} \end{pmatrix} \\ &= \begin{pmatrix} \sum_n y_n(1 - y_n)x_n^2 & \sum_n y_n(1 - y_n)x_n \\ \sum_n y_n(1 - y_n)x_n & \sum_n y_n(1 - y_n) \end{pmatrix}. \end{aligned} \tag{239}$$

Отметим, что для этой простой модели матрица Гессе не зависит от целевых значений, но является функцией параметров модели w и b, что соответствует тому факту, что функция ошибки не является квадратичной. Поскольку логистическая сигмоидная функция удовлетворяет условию $0 < \sigma(\cdot) < 1$, видно, что $y_n(1 - y_n)$ всегда является положительной величиной. Таким образом, получается, что элементы главной диагонали матрицы Гессе задаются суммой положительных слагаемых и, следовательно, сами являются положительными. Таким образом, след матрицы Гессе является положительным. Обратите внимание, что здесь игнорируется вырожденный случай, когда все точки данных идентичны, что приводит к следу, равному нулю, поскольку это не представляет практического интереса. Для определителя сначала зададим $c_n = y_n(1 - y_n)$, чтобы обозначения были более понятными.

Затем получаем

$$\begin{aligned}\det \mathbf{H} &= \left(\sum_n c_n x_n^2\right)\left(\sum_n c_n\right) - \left(\sum_n c_n x_n\right)^2 \\ &= \left(\sum_n c_n\right)\sum_n c_n \left\{ x_n - \frac{\left(\sum_n c_n x_n\right)}{\left(\sum_n c_n\right)} \right\}^2. \end{aligned} \tag{240}$$

Таким образом, определитель состоит из суммы слагаемых, каждое из которых является положительным, и, следовательно, сам по себе является положительным, а значит, и след, и определитель являются положительными. Поскольку определитель является произведением двух собственных значений матрицы Гессе, из этого следует, что либо оба собственных значения

являются положительными, либо оба являются отрицательными. Поскольку след является суммой собственных значений и при этом след также является положительным, из этого следует, что оба собственных значения должны быть положительными, а значит, матрица Гессе является положительно определенной.

7.6. Начнем с замены переменной, заданной формулой (7.10), которая позволяет записать функцию ошибки в виде (7.11). Установив значение функции ошибки $E(\mathbf{w})$ равным постоянной величине C, получаем

$$E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 = C.$$

Перегруппировав, получаем

$$\sum_i \lambda_i \alpha_i^2 = 2C - 2E(\mathbf{w}^\star) = \tilde{C},$$

где $\tilde{C}$ также является постоянной величиной. Это уравнение эллипса, оси которого совпадают с координатами, описываемыми переменными $\{\alpha_i\}$. Длина оси j определяется путем установки $\alpha_i = 0$ для всех $i \neq j$ и решения уравнения для α_j, что дает

$$\alpha_j = \left(\frac{\tilde{C}}{\lambda_j}\right)^{1/2},$$

которое обратно пропорционально квадратному корню соответствующего собственного значения.

7.7. Матрица $W \times W$ состоит из W^2 элементов. Если она симметрична, то элементы, не лежащие на главной диагонали, образуют пары равных значений. На диагонали находится W элементов, поэтому количество элементов, не лежащих на диагонали, равно $W^2 - W$, и только половина из них являются независимыми, что дает

$$\frac{W^2 - W}{2}$$

в качестве количества независимых элементов, не лежащих на диагонали. Если теперь добавить обратно W элементов на диагонали, то получим

$$\frac{W^2 - W}{2} + W = \frac{W(W+1)}{2}.$$

Наконец, добавим W элементов вектора градиента $\mathbf{b}$, чтобы получить

$$\frac{W(W+1)}{2} + W = \frac{W(W+1) + 2W}{2} = \frac{W^3 + 3W}{2} = \frac{W(W+3)}{2}.$$

7.8. Из свойства (2.52) гауссова распределения для одной точки данных получаем

$$\mathbb{E}[x_n] = \mu.$$

Две независимые точки данных не будут коррелированы, и, следовательно,

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2 \text{ при условии } n \neq m.$$

Таким образом, используя (2.53), получаем, что

$$\mathbb{E}[x_n x_m] = \delta_{nm}\sigma^2 + \mu^2.$$

Используя определение (7.65) $\bar{x}$, получаем

$$\mathbb{E}[\bar{x}] = \mu,$$

$$\begin{aligned}\mathbb{E}[\bar{x}^2] &= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[x_n x_m] \\ &= \frac{1}{N}\sigma^2 + \mu^2.\end{aligned}$$

Следовательно,

$$\begin{aligned}\mathbb{E}[(\bar{x} - \mu)^2] &= \mathbb{E}[\bar{x}^2 - 2\bar{x}\mu + \mu^2] \\ &= \frac{1}{N}\sigma^2 + \mu^2 - 2\mu^2 + \mu^2 = \frac{\sigma^2}{N}.\end{aligned}$$

7.9. Обратите внимание, что в оригинальном тексте в левой части уравнения (7.20) должно быть $\mathrm{var}[a_i^{(l)}]$ вместо $\mathrm{var}[z_i^{(l)}]$ (прим. авторов). Ожидаемое значение по $a_i^{(l)}$ включает усреднение как по распределению w_{ij}, так и по распределению $z_j^{(l-1)}$. Для заданного значения $z_j^{(l-1)}$ величина $w_i z_j^{(l-1)}$ имеет нулевое среднее значение, поскольку предполагается, что веса w_{ij} взяты из гауссова распределения $\mathcal{N}(0; \epsilon^2)$ с нулевым средним значением. Поскольку это верно для каждого значения j в сумме, получаем

$$\mathbb{E}[a_i^{(l)}] = 0. \tag{241}$$

Чтобы определить дисперсию $a_i^{(l)}$, примем во внимание, что величина $a_i^{(l)}$ состоит из суммы M слагаемых, которые сами по себе являются независимыми случайными переменными, и из (2.121) известен тот факт, что общая дисперсия суммы независимых переменных равна сумме дисперсий отдельных переменных. Таким образом, получаем

$$\begin{aligned}\mathrm{var}[a_i^{(l)}] &= M\,\mathrm{var}[w_{ij} z_j^{(l-1)}] \\ &= M\,\mathbb{E}_{w,z}[(w_{ij} z_j^{(l-1)})^2] - M\,\mathbb{E}_{w,z}[w_{ij} z_j^{(l-1)}]^2 \\ &= M\,\mathbb{E}_{w,z}[(w_{ij} z_j^{(l-1)})^2] \\ &= M\,\mathbb{E}_{w}[(w_{ij})^2]\mathbb{E}_{z}[(z_j^{(l-1)})^2],\end{aligned}$$

поскольку $\mathbb{E}_{w,z}[w_{ij}z_j^{(l-1)}] = 0$, как обсуждалось выше. Поскольку веса w_{ij} взяты из гауссова распределения $\mathcal{N}(0; \epsilon^2)$, получаем $\mathbb{E}[(w_{ij})^2] = \epsilon^2$. Чтобы найти $\mathbb{E}[(z_j^{(l-1)})^2]$, отметим, что

$$z_j^{(l-1)} = \text{ReLU}(a_j^{(l-1)})$$

и, следовательно,

$$(z_j^{(l-1)})^2 = \text{ReLU}(a_i^{(l-1)})^2.$$

При возведении в квадрат функции ReLU результат будет равен нулю, если аргумент меньше нуля или отрицателен, и будет равен квадрату аргумента, если аргумент положителен. Величина $a_i^{(l-1)}$ будет иметь симметричное распределение относительно 0, поскольку для каждого значения $z_j^{(l-1)}$ произведение $w_{ij}z_j^{(l-1)}$ будет иметь симметричное гауссово распределение, и, следовательно, общее распределение является суммой симметричных распределений, а значит, само по себе симметрично. Таким образом, при вычислении математического ожидания ReLU этой величины половина слагаемых будет равна нулю, а половина слагаемых будет равна квадрату аргумента и, следовательно,

$$\begin{aligned}\mathbb{E}[(z_j^{(l-1)})^2] &= \mathbb{E}[(\text{ReLU}(a_i^{(l-1)}))^2] \\ &= \frac{1}{2}\mathbb{E}[((a_i^{(l-1)})^2] \\ &= \frac{1}{2}\text{var}[(a_i^{(l-1)})^2] \\ &= \frac{1}{2}\lambda^2,\end{aligned}$$

где использовано свойство, что $a_i^{(l-1)}$ имеет нулевое среднее значение. Объединяя эти результаты, получаем

$$\text{var}[a_i^{(l)}] = \frac{M}{2}\epsilon^2\lambda^2,$$

как и требовалось. Наконец, если дисперсия $a_i^{(l-1)}$ также равна λ^2, тогда

$$\frac{M}{2}\epsilon^2\lambda^2 = \lambda^2,$$

из чего следует, что ϵ должно быть выбрано так, чтобы принимало значение

$$\epsilon = \sqrt{\frac{2}{M}}. \tag{242}$$

7.10. Взяв градиент от (7.7), получаем

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star). \tag{243}$$

Подставляя $(\mathbf{w} - \mathbf{w}^\star)$ с помощью (7.10), получаем

$$\nabla E = \mathbf{H}\Big(\sum_i \alpha_i \mathbf{u}_i\Big). \tag{244}$$

Используя (7.8), получаем

$$\nabla E = \sum_i \alpha_i \lambda_i \mathbf{u}_i. \tag{245}$$

Если рассматривать изменение $\Delta\alpha_i$ в коэффициентах α_i, то соответствующее изменение $\mathbf{w}$ получается путем взятия конечных разностей (7.10), что дает

$$\nabla \mathbf{w} = \sum_i \Delta\alpha_i \mathbf{u}_i. \tag{246}$$

Далее, используя (7.16), получаем

$$\Delta \mathbf{w} = -\eta \nabla E. \tag{247}$$

Подставляя это выражение в обе стороны, получаем

$$\sum_i \Delta\alpha_i \mathbf{u}_i = -\eta \sum_i \alpha_i \lambda_i \mathbf{u}_i. \tag{248}$$

Если умножить обе стороны на $\mathbf{u}_j^{\mathrm{T}}$ и использовать отношение ортонормированности (7.9), то в конечном итоге получаем

$$\Delta\alpha_j = -\eta \alpha_j \lambda_j, \tag{249}$$

как и требовалось.

7.11. Для малых значений μ можно провести разложение Тейлора первого члена правой части (7.34) по степеням μ, чтобы получить

$$\Delta \mathbf{w}^{(\tau-1)} = -\eta \nabla E(\mathbf{w}^{(\tau-1)}) - \eta\mu \nabla\nabla E(\mathbf{w}^{(\tau-1)}) \Delta \mathbf{w}^{(\tau-2)} + \mathcal{O}(\mu^2) + \mu \Delta \mathbf{w}^{(\tau-2)}.$$

Если теперь предположить, что $\eta = \mathcal{O}(\epsilon)$ и $\mu = \mathcal{O}(\epsilon)$, то можно пренебречь членами высшего порядка в разложении Тейлора, а также опустить член с коэффициентом $\eta\mu$, поскольку он равен $\mathcal{O}(\epsilon^2)$. Здесь было сделано допущение, что поверхность ошибки изменяется медленно и, следовательно, член матрицы Гессе $\nabla\nabla E$ равен $\mathcal{O}(1)$. Тогда получаем стандартную формулу для градиентного спуска с импульсом, определяемым формулой (7.31).

7.12. Если мы рекурсивно применим (7.66), то получим

$$\begin{aligned}\mu_n &= \beta\mu_{n-1} + (1-\beta)x_n \\ &= \beta(\beta\mu_{n-2} + (1-\beta)x_{n-1}) + (1-\beta)x_n \\ &= \beta^2\mu_{n-2} + \beta(1-\beta)x_{n-1} + (1-\beta)x_n \\ &= \beta^3\mu_{n-3} + \beta^2(1-\beta)x_{n-2} + \beta(1-\beta)x_{n-1} + (1-\beta)x_n \\ &= \ldots \\ &= \beta^n\mu_0 + \sum_{k=1}^{n}\beta^{k-1}(1-\beta)x_{n-k+1}.\end{aligned}$$

Теперь установим $\mu_0 = 0$, а затем возьмем математическое ожидание обеих сторон по отношению к распределению x, отметив, что $\{x_n\}$ являются независимыми, одинаково распределенными выборками из этого распределения. Это дает

$$\begin{aligned}\mathbb{E}[\mu_n] &= \sum_{k=1}^{n}\beta^{k-1}(1-\beta)\mathbb{E}[x_{n-k+1}] \\ &= \sum_{k=1}^{n}\beta^{k-1}(1-\beta)\bar{x},\end{aligned}$$

где $\bar{x}$ – это истинное среднее значение распределения x. Используя результат (7.67), это можно записать как

$$\mathbb{E}[\mu_n] = (1-\beta^n)\bar{x}.$$

Таким образом, получается, что $\mathbb{E}[\mu_n] \neq \bar{x}$ и, следовательно, оценка μ_n является смещенной. Это смещение легко исправить с помощью оценщика

$$\hat{\mu}_n = \frac{\mu}{1-\beta^n}, \tag{250}$$

который имеет свойство $\mathbb{E}[\mu_n] = \bar{x}$ и, следовательно, является несмещенным.

7.13. Приравнивая производную (7.69) по λ к нулю, получаем

$$\frac{\partial}{\partial\lambda}E(\mathbf{w}^{(\tau)} + \lambda\mathbf{d}) = 0. \tag{251}$$

Для вычисления этой производной можно использовать правило цепной производной. Определим

$$\mathbf{v} = \mathbf{w}^{(\tau)} + \lambda\mathbf{d}. \tag{252}$$

Тогда получаем

$$\begin{aligned}\frac{\partial}{\partial\lambda}E &= \sum_{i=1}^{M}\frac{\partial v_i}{\partial\lambda}\frac{\partial E}{\partial v_i}\\ &= \sum_{i=1}^{M}d_i\frac{\partial E}{\partial v_i}\\ &= \mathbf{d}^{\mathrm{T}}\nabla E,\end{aligned} \tag{253}$$

где $\{d_i\}$ – это компоненты $\mathbf{d}$. Эта производная исчезает для определенного значения $\lambda^\star$, которое затем определяет новое местоположение в весовом пространстве:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \lambda^\star\mathbf{d}. \tag{254}$$

Поэтому получаем

$$\mathbf{d}^{\mathrm{T}}\nabla E(\mathbf{w}^{(\tau)} + \lambda^\star\mathbf{d}) = \mathbf{d}^{\mathrm{T}}\nabla E(\mathbf{w}^{(\tau+1)}) = 0, \tag{255}$$

и, следовательно, градиент функции ошибки в минимуме линейного поиска ортогонален направлению поиска.

7.14. Суммируя обе части (7.50) по n для вычисления среднего значения выборки, получаем

$$\frac{1}{N}\sum_{n=1}^{N}\tilde{x}_{ni} = \frac{1}{N\sigma_i}\left(\sum_{n=1}^{N}x_{ni} - N\mu_i\right) = 0, \tag{256}$$

где использовано (7.48). Аналогично, если рассмотреть дисперсию выборки

$$\frac{1}{N}\sum_{n=1}^{N}(\tilde{x}_{ni} - 0)^2 = \frac{1}{N\sigma_i}\sum_{n=1}^{N}(x_{ni} - \mu_i)^2 = 1, \tag{257}$$

где используется (256) вместе с (7.49).

Глава 8. Обратное распространение

8.1. Из (8.12) получаем

$$\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j}, \tag{258}$$

что следует из правила цепочки вероятностей. Тогда из (8.8) получаем

$$\delta_k = \frac{\partial E_n}{\partial a_k}. \tag{259}$$

Вновь используя правило цепочки, получаем

$$\begin{aligned}\frac{\partial a_k}{\partial a_j} &= \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial a_j} \\ &= w_{kj}\frac{\partial z_j}{\partial a_j} \\ &= w_{kj}h'(a_j),\end{aligned} \tag{260}$$

где использованы (8.5) и (8.6). Подставляя эти результаты в (258), получаем

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k, \tag{261}$$

как и требовалось.

8.2. Уравнения прямого распространения в матричной форме задаются (6.19) в виде

$$\mathbf{z}^{(l)} = h^{(l)}\mathbf{W}^{(l)}\mathbf{z}^{(l-1)}, \tag{262}$$

где $\mathbf{W}^{(l)}$ – это матрица с элементами $\mathbf{w}_{jk}^{(l)}$, составляющими веса в сетевом слое l, а активационная функция $h^{(l)}(\cdot)$ действует на каждый элемент своего векторного аргумента независимо. Если определить $\boldsymbol{\delta}^{(l)}$ как вектор ошибок с элементами δ_j, то уравнения обратного распространения в матричной форме примут вид

$$\boldsymbol{\delta}^{(l-1)} = h^{(l-1)\prime}(\mathbf{a}^{(l-1)}) \odot \{(\mathbf{W}^{(l-1)})^{\mathrm{T}}\boldsymbol{\delta}^{(l)}\}, \tag{263}$$

где $\odot$ обозначает произведение Адамара, которое состоит из элементарного умножения двух векторов. Здесь следует отметить, что уравнение прямого распространения (8.5) включает суммирование по второму индексу w_{ji}, тогда как уравнение обратного распространения (8.13) включает суммирование по первому индексу. Следовательно, при записи уравнения обратного распространения в матричной форме используется транспонированная матрица, которая появляется в уравнении прямого распространения.

8.3. Здесь необходимо найти зависимость поправочного члена

$$\delta = E'(w_{ij}) - \frac{E(w_{ij}+\epsilon) - E(w_{ij}-\epsilon)}{2\epsilon} \tag{264}$$

от ϵ. С помощью разложения Тейлора можно переписать числитель первого члена (264) в виде

$$\begin{aligned}&E(w_{ij}) + \epsilon E'(w_{ij}) + \frac{\epsilon^2}{2}E''(w_{ij}) + \mathcal{O}(\epsilon^3) \\ &\quad - E(w_{ij}) + \epsilon E'(w_{ij}) - \frac{\epsilon^2}{2}E''(w_{ij}) + \mathcal{O}(\epsilon^3) = 2\epsilon E'(w_{ij}) + \mathcal{O}(\epsilon^3).\end{aligned}$$

Здесь следует отметить, что слагаемые ϵ^2 аннулируются. Подставляя это в (264), получаем

$$\delta = \frac{2\epsilon E'(w_{ij}) + \mathcal{O}(\epsilon^3)}{2\epsilon} - E'(w_{ij}) = \mathcal{O}(\epsilon^2). \tag{265}$$

8.4. Если ввести веса пропускаемого слоя **U** в модель, описанную в разделе 8.1.3, это повлияет только на последнее из уравнений прямого распространения (8.20), которое примет вид

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} u_{ki} x_i. \tag{266}$$

Здесь следует отметить, что нет необходимости включать входное смещение. Производная по u_{ki} может быть выражена с помощью выходного $\{\delta_k\}$ из (8.21),

$$\frac{\partial E}{\partial u_{ki}} = \delta_k x_i. \tag{267}$$

8.5. Альтернативная схема прямого распространения берет за отправную точку первую строку (8.29). Однако вместо того, чтобы продолжать с «рекурсивным» определением $\partial y_k / \partial a_j$, используем соответствующее определение для $\partial a_j / \partial x_i$. Выглядит это следующим образом:

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial x_i}, \tag{268}$$

где $\partial y_k / \partial a_j$ определяется формулой (8.33) для логистических сигмоидных выходных блоков, формулой (8.34) для выходных блоков softmax или просто как δ_{kj} для линейных выходных блоков. Определяем $\partial a_j / \partial x_i = w_{ji}$, если a_j находится в первом скрытом слое, и в противном случае

$$\frac{\partial a_j}{\partial x_i} = \sum_l \frac{\partial a_j}{\partial a_l} \frac{\partial a_l}{\partial x_i}, \tag{269}$$

где

$$\frac{\partial a_j}{\partial a_l} = w_{jl} h'(a_l). \tag{270}$$

Таким образом, можно оценить J_{ki} посредством прямого распространения $\partial a_j / \partial x_i$ с начальным значением w_{ij} наряду с a_j, используя (269) и (270).

8.6. Используя правило цепочки вместе с (8.5) и (8.77), получаем

$$\begin{aligned}\frac{\partial E_n}{\partial w_{kj}^{(2)}} &= \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}^{(2)}} \\ &= \delta_k z_j. \end{aligned} \tag{271}$$

Таким образом,

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\partial w_{k'j'}^{(2)}} = \frac{\partial \delta_k z_j}{\partial w_{k'j'}^{(2)}}, \tag{272}$$

и, поскольку z_j не зависит от весов второго слоя,

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\partial w_{k'j'}^{(2)}} &= z_j \frac{\partial \delta_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}}\frac{\partial a_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j z_{j'} M_{kk'},\end{aligned}$$

где вновь использовано правило цепочки вместе с (8.5) и (8.77). Если оба веса находятся в первом слое, снова используем правило цепочки, на этот раз вместе с (8.5), (8.12) и (8.13), чтобы получить

$$\begin{aligned}\frac{\partial E_n}{\partial w_{kj}^{(1)}} &= \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{kj}^{(1)}} \\ &= x_i \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= x_i h'(a_j)\sum_k w_{kj}^{(2)}\delta_k.\end{aligned}$$

Таким образом, получаем

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\partial w_{j'i'}^{(1)}} = \frac{\partial}{\partial w_{j'i'}}\left(x_i h'(a_j)\sum_k w_{kj}^{(2)}\delta_k\right). \tag{273}$$

Теперь примем во внимание, что x_i и $w_{kj}^{(2)}$ не зависят от $w_{j'i'}^{(1)}$, в то время как $h'(a_j)$ зависит только в случае, когда $j = j'$. Используя эти наблюдения вместе с (8.5), получаем

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\partial w_{j'i'}^{(1)}} = x_i x_{i'} h''(a_j) I_{jj'}\sum_k w_{kj}^{(2)}\delta_k + x_i h'(a_j)\sum_k w_{kj}^{(2)}\frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}}. \tag{274}$$

Из (8.5), (8.12), (8.13), (8.77) и правила цепочки получаем

$$\begin{aligned}\frac{\partial \delta_k}{\partial w^{(1)}_{j'i'}} &= \sum_{k'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial a_{j'}} \frac{\partial a_{j'}}{\partial w^{(1)}_{j'i'}} \\ &= x_{i'} h'(a_j) \sum_{k'} w^{(2)}_{k'j'} M_{kk'}. \end{aligned} \tag{275}$$

Подставляя это обратно в (274), получаем (**??**). Наконец, из (271) получаем

$$\frac{\partial^2 E_n}{\partial w^{(1)}_{ji} \partial w^{(2)}_{kj'}} = \frac{\partial \delta_k z_{j'}}{\partial w^{(1)}_{ji}}. \tag{276}$$

Используя (275), получаем

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w^{(1)}_{ij} \partial w^{(2)}_{kj'}} &= z_{j'} x_i h'(a_j) \sum_{k'} w^{(2)}_{k'j} M_{kk'} + \delta_k I_{jj'} h'(a_j) x_i \\ &= x_i h'(a_j) \left(\delta_k I_{jj'} + \sum_{k'} w^{(2)}_{k'j} M_{kk'} \right). \end{aligned}$$

8.7. Если ввести веса пропущенного слоя в модель, рассмотренную в разделе **??**, к трем случаям, уже рассмотренным в упражнении 8.6, добавляются три новых. Первую производную по весу пропущенного слоя u_{ki} можно записать следующим образом:

$$\frac{\partial E_n}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} x_i. \tag{277}$$

Используя это, можно рассмотреть первый новый случай, когда оба веса находятся в пропущенном слое

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial u_{k'i'}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial u_{k'i'}} x_i \\ &= M_{kk'} x_i x_{i'}, \end{aligned}$$

где также использовано (8.77). Когда один вес находится в пропущенном слое, а другой вес находится в слое между скрытым и выходным слоями, можно использовать (277), (8.5) и (8.77), чтобы получить

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial w^{(2)}_{k'j}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial w^{(2)}_{k'j}} x_i \\ &= M_{kk'} z_j x_i. \end{aligned}$$

Наконец, если один вес является весом пропущенного слоя, а другой находится в слое от входа к скрытому слою, тогда (277), (8.5), (8.12), (8.13) и (8.77) вместе дают

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki}\partial w_{ji'}^{(1)}} &= \frac{\partial}{\partial w_{ji'}^{(1)}}\left(\frac{\partial E_n}{\partial a_k}x_i\right)\\ &= \sum_{k'}\frac{\partial^2 E_n}{\partial a_k \partial a_{k'}}\frac{\partial a_{k'}}{\partial w_{ji'}^{(1)}}x_i\\ &= x_i x_{i'} h'(a_j)\sum_{k'} M_{kk'} w_{k'j}^{(2)}.\end{aligned}$$

8.8. Многомерная форма (8.38) имеет вид

$$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}(\mathbf{y}_n - \mathbf{t}_n). \tag{278}$$

Элементы первой и второй производных в этом случае принимают вид

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} \tag{279}$$

и

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N}\left\{\frac{\partial \mathbf{y}_n}{\partial w_j}^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} + (\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial^2 \mathbf{y}_n}{\partial w_i \partial w_j}\right\}. \tag{280}$$

Как и в одномерном случае, предположим, что второй член второй производной исчезает, и получаем

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n^{\mathrm{T}}, \tag{281}$$

где $\mathbf{B}_n$ – это матрица $W\times K$, а K – это размерность y_n с элементами

$$(\mathbf{B}_n)_{lk} = \frac{\partial y_{nk}}{\partial w_l}. \tag{282}$$

8.9. Взяв вторые производные (8.78) по двум весам w_r и w_s, получаем

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \sum_k \int\left\{\frac{\partial y_k}{\partial w_r}\frac{\partial y_k}{\partial w_s}\right\}p(\mathbf{x})\mathrm{d}\mathbf{x}\\ &+ \sum_k \int\left\{\frac{\partial^2 y_k}{\partial w_r \partial w_s}(y_k(\mathbf{x}) - \mathbb{E}_{t_k}[t_k|\mathbf{x}])\right\}p(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned} \tag{283}$$

Используя результат (4.37), согласно которому выходы $y_k(\mathbf{x})$ обученной сети представляют собой условные средние значения целевых данных, становится понятно, что второй член в (283) исчезает. Таким образом, матрица Гессе задается интегралом членов, содержащих только произведения первых производных. Для конечного набора данных этот результат можно записать в виде

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k}\frac{\partial y_k^n}{\partial w_r}\frac{\partial y_k^n}{\partial w_s}, \tag{284}$$

что идентично (8.40) с точностью до коэффициента масштабирования.

8.10. Если взять градиент (6.33) по $\mathbf{w}$, то получим

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\frac{\partial E}{\partial a_n}\nabla a_n = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n,$$

где используем результат, доказанный ранее в решении упражнения 6.14. Взяв вторые производные, получаем

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\left\{\frac{\partial y_n}{\partial a_n}\nabla a_n \nabla a_n + (y_n - t_n)\nabla\nabla a_n\right\}.$$

Опуская последний член и используя результат (5.72) для производной логистической сигмоидной функции, доказанный в решении упражнения 5.18, в конечном итоге получаем

$$\nabla\nabla E(\mathbf{w}) \simeq \sum_{n=1}^{N} y_n(1 - y_n)\nabla a_n \nabla a_n = \sum_{n=1}^{N} y_n(1 - y_n)\mathbf{b}_n\mathbf{b}_n^{\mathrm{T}},$$

где $\mathbf{b}_n \equiv \nabla a_n$.

8.11. Используя правило цепной производной, первую производную (6.36) можно записать в виде

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\partial E}{\partial a_{nk}}\frac{\partial a_{nk}}{\partial w_i}. \tag{285}$$

Из упражнения 6.15 известно, что

$$\frac{\partial E}{\partial a_{nk}} = y_{nk} - t_{nk}. \tag{286}$$

Используя это и (5.78), можно получить производную (285) по w_j в виде

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl})\frac{\partial a_{nk}}{\partial w_i}\frac{\partial a_{nl}}{\partial w_j} + (y_{nk} - t_{nk})\frac{\partial^2 a_{nk}}{\partial w_i \partial w_j}\right).$$

Для обученной модели выходы сети будут приближаться к условным вероятностям классов, и поэтому последний член в скобках исчезнет в пределе большого набора данных, в результате чего получаем

$$(\mathbf{H})_{ij} = \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl})\frac{\partial a_{nk}}{\partial w_i}\frac{\partial a_{nl}}{\partial w_j}. \tag{287}$$

8.12. Предположим, что у нас уже есть обратная матрица Гессе, полученная с помощью первых L точек данных. Отделив вклад точки данных $L + 1$ в (8.40), получаем

$$\mathbf{H}_{L+1} = \mathbf{H}_L + \nabla a_{L+1} \nabla a_{L+1}^{\mathrm{T}}. \tag{288}$$

Далее рассмотрим матричную тождественность (8.80). Если теперь тождественно сопоставить $\mathbf{H}_L$ с $\mathbf{M}$, а $\mathbf{b}_{L+1}$ с $\mathbf{v}$, то получим

$$\mathbf{H}_{L+1}^{-1} = \mathbf{H}_L^{-1} - \frac{\mathbf{H}_L^{-1} \nabla a_{L+1} \nabla a_{L+1}^{\mathrm{T}} \mathbf{H}_L^{-1}}{1 + \nabla a_{L+1}^{\mathrm{T}} \mathbf{H}_L^{-1} \nabla a_{L+1}}. \tag{289}$$

Таким образом, точки данных последовательно поглощаются, пока $L+1 = N$ и весь набор данных не будет обработан. Таким образом, этот результат представляет собой процедуру вычисления обратной матрицы Гессе с помощью одного прохода по набору данных. Начальная матрица $\mathbf{H}_0$ выбирается равной $\alpha\mathbf{I}$, где α – это небольшое количество, так что алгоритм фактически находит обратную матрицу $\mathbf{H} + \alpha\mathbf{I}$. Результаты при этом не особенно чувствительны к точному значению α.

8.13. Функция $h(\cdot)$ задается с помощью soft ReLU:

$$h(a) = \ln(1 + \exp(a)), \tag{290}$$

а ее производная задается следующим образом:

$$h'(a) = \frac{\exp(a)}{1 + \exp(a)}. \tag{291}$$

Теперь можно использовать правило цепной производной в виде

$$\frac{\mathrm{d}z}{\mathrm{d}w_1} = \frac{\exp(w_1 x + b_1)}{1 + \exp(w_1 x + b_1)} x. \tag{292}$$

Используя (8.44) и (291), получаем

$$\frac{\mathrm{d}y}{\mathrm{d}z} = \frac{\exp(w_2 x + b_2)}{1 + \exp(w_2 x + b_2)} w_2. \tag{293}$$

Аналогично, используя (8.45) и (291), получаем

$$\frac{\mathrm{d}y}{\mathrm{d}w_1} = \frac{w_2 x \exp(w_1 x + b_1 + b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)])}{(1 + \exp(w_1 x + b_1))\big(1 + \exp(b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)])\big)}. \tag{294}$$

Наконец, объединив эти производные с помощью правила цепной производной, а затем подставив z с помощью (8.44), получаем

$$\frac{\mathrm{d}y}{\mathrm{d}w_1} = \frac{w_2 x \exp(w_1 x + b_1 + b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)])}{(1 + \exp(w_1 x + b_1))\big(1 + \exp(b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)])\big)}. \tag{295}$$

8.14. Уравнения следа оценки задаются непосредственно из определения логистического отображения:

$$L_1 = x, \tag{296}$$

$$L_2 = 4L_1(1 - L_1), \tag{297}$$

$$L_3 = 4L_2(1 - L_2), \tag{298}$$

$$L_4 = 4L_3(1 - L_3). \tag{299}$$

$$\tag{300}$$

Соответствующие явные функции без упрощения задаются следующим образом:

$$L_1(x) = x, \tag{301}$$

$$L_2(x) = 4x(1 - x), \tag{302}$$

$$L_3(x) = 16x(1 - x)(1 - 2x)^2, \tag{303}$$

$$L_4(x) = 64x(1 - x)(1 - 2x)^2(1 - 8x + 8x^2)^2. \tag{304}$$

Наконец, взяв производные, получаем следующие выражения опять же без упрощения:

$$L_1'(x) = 1, \tag{305}$$

$$L_2'(x) = 4(1 - x) - 4x, \tag{306}$$

$$L_3'(x) = 16(1 - x)(1 - 2x)^2 - 16x(1 - 2x)^2 - 64x(1 - x)(1 - 2x), \tag{307}$$

$$\begin{aligned} L_4'(x) = {} & 128x(1 - x)(-8 + 16x)(1 - 2x)^2(1 - 8x + 8x^2) \\ & + 64(1 - x)(1 - 2x)^2(1 - 8x + 8x^2)^2 \\ & - 64x(1 - 2x)^2(1 - 8x + 8x^2)^2 \\ & - 256x(1 - x)(1 - 2x)(1 - 8x + 8x^2)^2. \end{aligned} \tag{308}$$

Здесь следует отметить, что сложность выражений для производных растет гораздо быстрее, чем сложность выражений для соответствующих функций.

8.15. Для получения уравнений прямого режима применим (8.57) в виде

$$\dot{v}_i = \sum_{j \in \mathrm{pa}(i)} \dot{v}_j \frac{\partial v_i}{\partial v_j},$$

где pa(i) обозначает родителей узла i в диаграмме трассировки оценки. Используя диаграмму трассировки оценки на рис. 8.4 вместе с (8.50)–(8.56), получаем

$$\dot{v}_1 = 1, \tag{309}$$

$$\dot{v}_2 = 0, \tag{310}$$

$$\dot{v}_3 = \dot{v}_1 \frac{\partial v_3}{\partial v_1} + \dot{v}_2 \frac{\partial v_3}{\partial v_2} = \dot{v}_1 v_2 + \dot{v}_2 v_1, \tag{311}$$

$$\dot{v}_4 = \dot{v}_2 \frac{\partial v_4}{\partial v_2} = \dot{v}_2 \cos(v_2), \tag{312}$$

$$\dot{v}_5 = \dot{v}_3 \frac{\partial v_5}{\partial v_3} = \dot{v}_3 \exp(v_3), \tag{313}$$

$$\dot{v}_6 = \dot{v}_3 \frac{\partial v_6}{\partial v_3} + \dot{v}_4 \frac{\partial v_6}{\partial v_4} = \dot{v}_3 - \dot{v}_4, \tag{314}$$

$$\dot{v}_7 = \dot{v}_5 \frac{\partial v_7}{\partial v_5} + \dot{v}_6 \frac{\partial v_7}{\partial v_6} = \dot{v}_5 + \dot{v}_6. \tag{315}$$

8.16. Для получения уравнений обратного режима применяем (8.69) в виде

$$\bar{v}_i = \sum_{j \in \mathrm{ch}(i)} \bar{v}_j \frac{\partial v_j}{\partial v_i}. \tag{316}$$

Здесь ch(i) обозначает дочерние элементы узла i в графе трассировки оценки. Используя диаграмму трассировки оценки на рис. 8.4 вместе с (8.50)–(8.56) и начиная с выхода графа и двигаясь в обратном направлении, получаем

$$\bar{v}_7 = 1, \tag{317}$$

$$\bar{v}_6 = \bar{v}_7 \frac{\partial v_7}{\partial v_6} = \bar{v}_7, \tag{317}$$

$$\bar{v}_5 = \bar{v}_7 \frac{\partial v_7}{\partial v_5} = \bar{v}_7, \tag{317}$$

$$\bar{v}_4 = \bar{v}_6 \frac{\partial v_6}{\partial v_4} = -\bar{v}_6, \tag{317}$$

$$\bar{v}_3 = \bar{v}_5 \frac{\partial v_5}{\partial v_3} + \bar{v}_6 \frac{\partial v_6}{\partial v_3} = \bar{v}_5 v_5 + \bar{v}_6, \tag{317}$$

$$\bar{v}_2 = \bar{v}_3 \frac{\partial v_3}{\partial v_2} + \bar{v}_4 \frac{\partial v_4}{\partial v_2} = \bar{v}_3 v_1 + \bar{v}_4 \cos(v_2), \tag{317}$$

$$\bar{v}_1 = \bar{v}_3 \frac{\partial v_3}{\partial v_1} = \bar{v}_3 v_2. \tag{317}$$

8.17. Из (8.49) получаем следующее выражение для частной производной

$$\frac{\partial f}{\partial x_1} = x_2 + x_2 \exp(x_1 x_2). \tag{324}$$

Вычисляя это для $(x_1 x_2) = (1; 2)$, получаем

$$\left.\frac{\partial f}{\partial x_1}\right|_{x_1=1,\, x_2=2} = 2 + 2\exp(2). \tag{325}$$

Из уравнений трассировки вычислений (8.50)–(8.56) получаем

$$v_1 = 1, \tag{326}$$

$$v_2 = 2, \tag{327}$$

$$v_3 = 2, \tag{328}$$

$$v_4 = \sin(2), \tag{329}$$

$$v_5 = \exp(2), \tag{330}$$

$$v_6 = 2 - \sin(2), \tag{331}$$

$$v_7 = 2 + \exp(2) - \sin(2). \tag{332}$$

Для тангенциальных переменных можно использовать уравнения от (8.58) до (8.64), чтобы получить

$$\dot{v}_1 = 1, \tag{333}$$

$$\dot{v}_2 = 0, \tag{334}$$

$$\dot{v}_3 = 2, \tag{335}$$

$$\dot{v}_4 = 0, \tag{336}$$

$$\dot{v}_5 = 2\exp(2), \tag{337}$$

$$\dot{v}_6 = 2, \tag{338}$$

$$\dot{v}_7 = 2 + 2\exp(2), \tag{339}$$

и, таким образом, получается, что $\dot{v}_7$ действительно представляет собой правильное значение производной, заданной формулой (325). Аналогично можно использовать уравнения трассировки оценки обратного режима ав-

томатического дифференцирования (8.70)–(8.76) для оценки сопряженных переменных следующим образом:

$$\bar{v}_7 = 1, \tag{340}$$

$$\bar{v}_6 = 1, \tag{341}$$

$$\bar{v}_5 = 1, \tag{342}$$

$$\bar{v}_4 = -1, \tag{343}$$

$$\bar{v}_3 = \exp(2) + 1, \tag{344}$$

$$\bar{v}_2 = (\exp(2) + 1) - \cos(2), \tag{345}$$

$$\bar{v}_1 = 2\exp(2) + 2. \tag{346}$$

Из (8.68) получаем

$$\bar{v}_1 = \frac{\partial f}{\partial x_1}, \tag{347}$$

и снова видно, что это согласуется с требуемой производной.

8.18. Векторы $\mathbf{e}_1, \ldots, \mathbf{e}_D$ образуют полный ортонормированный базис, поэтому можно разложить произвольный D-мерный вектор $\mathbf{r}$ в виде

$$\mathbf{r} = \sum_{i=1}^{D} \alpha_i \mathbf{e}_i. \tag{348}$$

Умножив обе части на $\mathbf{e}_j^{\mathrm{T}}$, получаем

$$r_j = \mathbf{e}_j^{\mathrm{T}} \mathbf{r} = \alpha_j, \tag{349}$$

где r_j – j-я компонента $\mathbf{r}$. Таким образом, разложение можно записать в виде

$$\mathbf{r} = \sum_{i=1}^{D} r_i \mathbf{e}_i. \tag{350}$$

Умножив обе части на матрицу Якоби, получаем

$$\mathbf{J}\mathbf{r} = \sum_{i=1}^{D} r_i \mathbf{J}\mathbf{e}_i = \sum_{i=1}^{D} r_i \frac{\partial \mathbf{f}}{\partial x_i}, \tag{351}$$

где $\mathbf{f}(\mathbf{x})$ – исходная сетевая функция с элементами $f_k(\mathbf{x})$. Это можно интерпретировать как один проход автоматического дифференцирования в прямом режиме, в котором тангенциальные переменные, связанные с входными переменными, задаются как $\dot{x}_i = r_i$.

Чтобы увидеть это более наглядно, можно ввести функцию $\mathbf{g}(z)$, где z – это скалярная переменная, а элементы $\mathbf{g}$ задаются как $g_i(z) = r_i z$. С точки

зрения диаграммы сети это можно рассматривать как введение дополнительного слоя от одного входа z к исходным входам $\{x_i\}$. Общая составная функция может быть записана как $\mathbf{f}(\mathbf{g}(z))$, которая теперь является функцией с одним входом, поэтому ее матрица Якоби является матрицей с одним столбцом, которая может быть вычислена за один проход автоматического дифференцирования в прямом режиме. Элементы этого вектора задаются формулой

$$\frac{\partial f_k}{\partial z} = \sum_{i=1}^{D} \frac{\partial f_k}{\partial x_i} \frac{\partial x_i}{\partial z} = \sum_{i=1}^{D} J_{ki} r_i = (\mathbf{Jr})_k \tag{352}$$

и, следовательно, являются элементами произведения вектора Якоби, как и требуется. Тангенциальные переменные на входах в основную сеть тогда задаются формулой

$$\dot{x}_i = \frac{\partial x_i}{\partial z} = r_i. \tag{353}$$

Таким образом, получается, что если тангенциальная переменная $\dot{x}_i$ для каждого входа i задана как соответствующий элемент r_i массива $\mathbf{r}$, то требуемое произведение вектора Якоби будет вычислено за один проход автоматического дифференцирования в прямом режиме.

Глава 9. Регуляризация

9.1. Начнем с доказательства того, что совокупность поворотов на кратные 90° образует группу:

- **замкнутость**: предположим, что $\mathcal{A}$ представляет поворот на $a°$, а $\mathcal{B}$ – поворот на $b°$ и что a и b являются кратными 90°, т. е. $\mathcal{A}$ и $\mathcal{B}$ являются элементами множества. Следовательно, $\mathcal{A} \circ \mathcal{B}$ представляет поворот на $a + b$, который также является кратным 90, поэтому этот новый поворот также является элементом множества;
- **ассоциативность**: теперь предположим, что есть три поворота, $\mathcal{A}$, $\mathcal{B}$ и $\mathcal{C}$, которые представляют повороты на $a°$, $b°$ и $c°$ соответственно, опять же каждый из которых кратен 90. Теперь видно, что как $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}$, так и $\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})$ соответствуют поворотам на $(a + b + c)$, что делает множество ассоциативным при составлении поворотов;
- **тождественность**: поворот на 0° входит в набор и также не изменяет другие повороты при их композиции;
- **инверсия**: если взять элемент $\mathcal{A}$ в множестве, которое, в свою очередь, представляет поворот на $a°$, то его инверсия $\mathcal{A}^{-1}$ будет поворотом на $360 - a$, что означает, что $\mathcal{A} \circ \mathcal{A}^{-1}$ даст поворот на 360°, который совпадает с поворотом на 0°, который является тождеством.

Подтверждение выполнения всех этих четырех аксиом является доказательством того, что это действительно группа. Теперь сделаем то же самое для группы переносов объекта в двумерной плоскости.

- **Замкнутость**: предположим, что $\mathcal{A}$ представляет собой перенос a_x в x и a_y в y, а $\mathcal{B}$ представляет собой перенос b_x на x и b_y в y. Таким образом, $\mathcal{A} \circ \mathcal{B}$ представляет собой перенос на величину $a_x + b_x$ в x и $a_y + b_y$ в y, что также является переносом в двумерной плоскости.
- **Ассоциативность**: предположим, что есть три переноса, $\mathcal{A}$, $\mathcal{B}$ и $\mathcal{C}$, которые соответствуют переносам a_x, b_x и c_x в x и a_y, b_y и c_y в y. Получается, что как $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}$, так и $\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})$ соответствуют переносам a_x, b_x и c_x в x и a_y, b_y и c_y в y.
- **Тождественность**: комбинация переноса 0 в обоих измерениях с любым другим переносом оставит последний без изменений, и поэтому перенос 0 в обоих измерениях является тождественностью для этой группы.
- **Инверсия**: если взять элемент $\mathcal{A}$ в множестве, которое вновь представляет перенос a_x в x и a_y в y, можно увидеть, что композиция этого с переносом $-a_x$ в x и $-a_y$ в y дает нам перенос 0 в обоих измерениях, а это означает, что этот отрицательный перенос является инверсией $\mathcal{A}$.

9.2. Пусть

$$\begin{aligned}\tilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) \\ &= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni},\end{aligned}$$

где $y_n = y(x_n, \mathbf{w})$ и $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$, при этом применяется (9.52). Тогда на основании (9.53) определяем:

$$\begin{aligned}\tilde{E} &= \frac{1}{2}\sum_{n=1}^{N}\{\tilde{y}_n - t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\{\tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2\} \\ &= \frac{1}{2}\sum_{n=1}^{N}\left\{\tilde{y}_n^2 + 2\tilde{y}_n \sum_{i=1}^{D} w_i \epsilon_{ni} + \left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2 \right. \\ &\qquad \left. - 2t_n y_n - 2t_n \sum_{i=1}^{D} w_i \epsilon_{ni} + t_n^2\right\}.\end{aligned}$$

Если взять математическое ожидание $\tilde{E}$ при распределении ϵ_{ni}, то можно увидеть, что второй и пятый члены исчезают, поскольку $\mathbb{E}[\epsilon_{ni}] = 0$, а для третьего члена получаем

$$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2\right] = \sum_{i=1}^{D} w_i^2 \sigma^2,$$

поскольку все ϵ_{ni} независимы с дисперсией σ^2.

Из этого и (9.53) получаем, что

$$\mathbb{E}[\tilde{E}] = E_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2 \sigma^2,$$

как и требовалось.

9.3. Сначала запишем формулу градиентного спуска в непрерывном времени t в виде

$$\mathbf{w}(t+\epsilon) = \mathbf{w}(t) - \epsilon\tilde{\eta}\nabla\Omega(\mathbf{w}), \tag{354}$$

где ϵ представляет собой некоторый конечный временной шаг, при этом есть определение $\tilde{\eta} = \eta/\epsilon$. Теперь проводим разложение Тейлора левой части по степеням ϵ, чтобы получить

$$\mathbf{w}(t) + \frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t}\epsilon + \mathcal{O}(\epsilon^2) = \mathbf{w}(t) - \epsilon\tilde{\eta}\nabla\Omega(\mathbf{w}). \tag{355}$$

Теперь вычисляем предел $\epsilon \to 0$, чтобы получить

$$\frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t} = -\tilde{\eta}\nabla\Omega(\mathbf{w}). \tag{356}$$

Далее подставляем $\Omega(\mathbf{w})$ с использованием

$$\Omega(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}, \tag{357}$$

что дает

$$\frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t} = -\tilde{\eta}\mathbf{w}. \tag{358}$$

У этого уравнения есть решение:

$$\mathbf{w}(t) = \mathbf{w}(0)\exp\{-\tilde{\eta}t\}, \tag{359}$$

что легко проверить путем подстановки. Таким образом, элементы $\mathbf{w}$ экспоненциально затухают до нуля.

9.4. С преобразованными входами, весами и смещениями (9.6) принимает вид

$$z_j = h\left(\sum_i \tilde{w}_{ji}\tilde{x}_i + \tilde{w}_{j0}\right).$$

С помощью (9.8)–(9.10) можно переписать аргумент $h(\cdot)$ в правой части как

$$\begin{aligned}\sum_i \frac{1}{a} w_{ji}(ax_i + b) + w_{j0} - \frac{b}{a}\sum_i w_{ji} \\ = \sum_i w_{ji} x_i + \frac{b}{a}\sum_i w_{ji} + w_{j0} - \frac{b}{a}\sum_i w_{ji} \\ = \sum_i w_{ji} x_i + w_{j0}.\end{aligned}$$

Точно так же, с преобразованными выходами, весами и смещениями, (9.7) приобретает вид

$$\tilde{y}_k = \sum_i \tilde{w}_{kj} z_j + \tilde{w}_{k0}.$$

Используя (9.11)–(9.13), можно переписать это выражение в виде

$$\begin{aligned}cy_k + d &= \sum_k c w_{kj} z_j + c w_{k0} + d \\ &= c\left(\sum_i w_{kj} z_j + w_{k0}\right) + d.\end{aligned}$$

Вычитая d и затем деля на c обе стороны, возвращаемся к исходному виду (9.7).

9.5. Ограничение (9.20) можно переписать в виде

$$\sum_{j=1}^{M} |w_j|^q - \eta \le 0. \tag{360}$$

Это ограничение может быть выполнено путем добавления слагаемого к нерегуляризованной ошибке $E(\mathbf{w})$ с использованием множителя Лагранжа, который обозначим как $\lambda/2$, что дает

$$E(\mathbf{w}) + \frac{\lambda}{2}\left(\sum_{j=1}^{M} |w_j|^q - \eta\right). \tag{361}$$

Поскольку член $\lambda\eta/2$ является константой относительно $\mathbf{w}$, минимизация (361) эквивалентна минимизации

$$E(\mathbf{w}) + \frac{\lambda}{2}\sum_{j=1}^{M} |w_j|^q. \tag{362}$$

Если ограничение активно, то $\lambda \neq 0$ (см. приложение C) и, следовательно,

$$\sum_{j=1}^{M} |w_j|^q = \eta. \tag{363}$$

Сила регуляризации увеличивается с увеличением λ, приводя веса к меньшим значениям. Точно так же сила ограничения увеличивается с уменьшением η, снова приводя веса к меньшим значениям. Однако точная зависимость между λ и η зависит от формы $E(\mathbf{w})$.

9.6. Градиент (9.56) задается как

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star),$$

и, следовательно, формула обновления (9.57) принимает вид

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star).$$

Умножив обе части на $\mathbf{u}_j^{\mathrm{T}}$, получаем

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (364)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathrm{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j u_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^\star), && (365) \end{aligned}$$

где использовано (9.59). Чтобы показать, что

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\}w_j^\star$$

для $\tau = 1, 2, \ldots$, можно использовать доказательство по индукции. Для $\tau = 1$ вспомним, что $\mathbf{w}^{(0)} = \mathbf{0}$, и подставим это в (365), что даст

$$\begin{aligned} w_j^{(1)} &= w_j^{(0)} - \rho\eta_j(w_j^{(0)} - w_j^\star) \\ &= \rho\eta_j w_j^\star \\ &= \{1 - (1 - \rho\eta_j)\}w_j^\star. \end{aligned}$$

Теперь предположим, что результат справедлив для $\tau = N - 1$, и затем воспользуемся (365):

$$\begin{aligned} w_j^{(N)} &= w_j^{(N-1)} - \rho\eta_j(w_j^{(N-1)} - w_j^\star) \\ &= w_j^{(N-1)}(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \{1 - (1 - \rho\eta_j)^{N-1}\}w_j^\star(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \{(1 - \rho\eta_j) - (1 - \rho\eta_j)^N\}w_j^\star + \rho\eta_j w_j^\star \\ &= \{1 - (1 - \rho\eta_j)^N w_j^\star, \end{aligned}$$

как и требовалось. При условии, что $|1 - \rho\eta_j| < 1$, имеем $(1 - \rho\eta_j)^\tau \to 0$, поскольку $\tau \to \infty$, и, следовательно, $\{1 - (1 - \rho\eta_j)^N \to 1$ и $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$. Если τ конечен, но при этом $\eta_j \gg (\rho\tau)^{-1}$, τ все равно должен быть большим, поскольку $\eta_j\rho\tau \gg 1$, даже если $|1 - \rho\eta_j| < 1$. Если τ велик, из приведенного выше рассуждения следует, что $w_j^{(\tau)} \simeq w_j^\star$. Если, с другой стороны, $\eta_j \ll (\rho\tau)^{-1}$, это означает, что $\rho\,\eta_j$ должно

быть малым, поскольку $\rho\eta_j\tau \ll 1$ и τ является целым числом, большим или равным единице. Если разложить

$$(1-\rho\eta_j)^\tau = 1 - \tau\rho\eta_j + O(\rho\eta_j^2)$$

и вставить это в (9.58), то получим

$$\begin{aligned} |w_j^{(\tau)}| &= |\{1-(1-\rho\eta_j)^\tau\}w_j^\star| \\ &= |\{1-(1-\tau\rho\eta_j + O(\rho\eta_j^2))\}w_j^\star| \\ &\simeq \tau\rho\eta_j|w_j^\star| \ll |w_j^\star|. \end{aligned}$$

9.7. Предположим, что набор весов $w_1, \ldots, w_K$ является общим, так что $w_1 = w_2 = \ldots = w_K = \lambda$. Производную от функции ошибки по λ можно вычислить с помощью правила цепной производной

$$\frac{\partial E}{\partial \lambda} = \sum_{i=1}^{K} \frac{\partial E}{\partial w_i}\frac{\partial w_i}{\partial \lambda} = \sum_{i=1}^{K}\frac{\partial E}{\partial w_i}, \tag{366}$$

где используется

$$\frac{\partial w_i}{\partial \lambda} = 1. \tag{367}$$

Следовательно, сначала необходимо запустить стандартный алгоритм обратного распространения (или автоматическое дифференцирование) для оценки индивидуальных градиентов $\partial E/\partial w_i$ для всех весов. Затем для каждой группы весов в сети, которые являются общими, необходимо суммировать градиенты по всем этим весам и использовать полученный комбинированный градиент для обновления весов в этой группе. Обратите внимание, что, пока веса в каждой группе инициализированы с одинаковым значением, это гарантирует, что они останутся равными после обновления.

9.8. Из формулы

$$p(w) = \sum_{j=1}^{M}\pi_j \mathcal{N}(w|\mu_j, \sigma_j^2) \tag{368}$$

можно определить следующие вероятности:

$$p(j) = \pi_j, \tag{369}$$

$$p(w|j) = \mathcal{N}(w|\mu_j, \sigma_j^2). \tag{370}$$

Следовательно, из теоремы Байеса получаем

$$p(j|w) = \frac{p(j)p(w|j)}{p(w)} = \frac{\pi_j \mathcal{N}(w|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w|\mu_k, \sigma_k^2)}. \tag{371}$$

9.9. Этот результат легко проверить, взяв производную от (9.22) с применением (2.49) и стандартных производных, что дает

$$\frac{\partial \Omega}{\partial w_i} = \frac{1}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)} \sum_j \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \frac{(w_i - \mu_j)}{\sigma^2}.$$

Объединяя это с (9.23) и (9.24), сразу получаем второй член (9.25).

9.10. Поскольку μ_js появляются только в членах регуляризации $\Omega(\mathbf{w})$, из (9.23) получаем

$$\frac{\partial \tilde{E}}{\partial \mu_j} = \lambda \frac{\partial \Omega}{\partial \mu_j}. \tag{372}$$

Используя (3.25), (9.22) и (9.24) и стандартные правила дифференцирования, вычислим производную $\Omega(\mathbf{w})$ следующим образом:

$$\begin{aligned}\frac{\partial \tilde{E}}{\partial \mu_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}(w_i|\mu_{j'}, \sigma_{j'}^2)} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \frac{w_i - \mu_j}{\sigma_j^2} \\ &= -\sum_i \gamma_j(\mathbf{w}_i) \frac{w_i - \mu_j}{\sigma_j^2}.\end{aligned}$$

Объединяя это с (372), получаем (9.26).

9.11. Следуя той же линии рассуждений, что и в решении 9.10, нужно определить производную $\Omega(\mathbf{w})$ по σ_j. Снова используя (3.25), (9.22) и (9.24) и стандартные правила дифференцирования, находим, что она равна

$$\begin{aligned}\frac{\partial \tilde{E}}{\partial \sigma_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}(w_i|\mu_{j'}, \sigma_{j'}^2)} \pi_j \frac{1}{(2\pi)^{1/2}} \left\{ -\frac{1}{\sigma_j^2} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right) \right. \\ &\quad \left. + \frac{1}{\sigma_j} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right) \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\} \\ &= \sum_i \gamma_j(w_i) \left\{ \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\}.\end{aligned}$$

Объединяя это с (372), получаем (9.28).

9.12. Из определения (9.30) имеем

$$\pi_k = \frac{\exp(\eta_k)}{\sum_l \exp(\eta_l)}. \tag{373}$$

Взятие производной дает

$$\begin{aligned}\frac{\partial \pi_k}{\partial \eta_j} &= \frac{\exp(\eta_k)}{\sum_l \exp(\eta_l)}\delta_{jk} - \frac{\exp(\eta_k)}{\left(\sum_l \exp(\eta_l)\right)^2}\exp(\eta_j) \\ &= \delta_{jk}\pi_k - \pi_j\pi_k. \end{aligned} \tag{374}$$

Из (9.22) и (**??**) получаем

$$\begin{aligned}\frac{\partial \tilde{E}}{\partial \eta_j} &= -\lambda \sum_i \sum_k \gamma_k(w_i)\frac{1}{\pi_k}\frac{\partial \pi_k}{\partial \eta_j} \\ &= -\lambda \sum_i \sum_k \gamma_k(w_i)\{\delta_{jk} - \pi_j\} \\ &= \lambda \sum_i \{\pi_j - \gamma_j(w_i)\}, \end{aligned} \tag{375}$$

где использован тот факт, что $\sum_k \gamma_k(w_i) = 1$ для всех i.

9.13. Результат легко доказать, подставляя (9.36) в (9.37), а затем подставляя (9.35) в полученное выражение, что дает

$$\begin{aligned}\mathbf{y} &= \mathbf{F}_3(\mathbf{z}_2) + \mathbf{z}_2 \\ &= \mathbf{F}_3(\mathbf{F}_2(\mathbf{z}_1) + \mathbf{z}_1) + \mathbf{F}_2(\mathbf{z}_1) + \mathbf{z}_1 \\ &= \mathbf{F}_3(\mathbf{F}_2(\mathbf{F}_1(\mathbf{x}) + \mathbf{x}) + \mathbf{F}_1(\mathbf{x}) + \mathbf{x}) \\ &\quad + \mathbf{F}_2(\mathbf{F}_1(\mathbf{x}) + \mathbf{x})) + \mathbf{F}_1(\mathbf{x}) + \mathbf{x}. \end{aligned} \tag{376}$$

9.14. Используя (9.49), мы можем переписать (9.47) как

$$\begin{aligned}E_{\text{COM}} &= \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}\left[\left\{\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] = \frac{1}{M}E_{\text{AV}}, \end{aligned}$$

где на последнем шаге использовано (9.46).

9.15. Начнем с перегруппировки правой части (9.46), перенеся множитель $1/M$ внутрь суммы, а оператор математического ожидания – за пределы суммы, что дает

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2\right].$$

Если затем отождествить $\epsilon_m(\mathbf{x})$ и $1/M$ с x_i и λ_i в (2.102) соответственно и взять $f(x) = x^2$, то из (2.102) следует, что

$$\left(\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\right)^2 \leq \sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2.$$

Поскольку это верно для всех значений $\mathbf{x}$, оно должно быть верно и для математического ожидания по $\mathbf{x}$, что и является доказательством (9.64).

9.16. Если функция $E(y(\mathbf{x}))$ является выпуклой, можно применить (2.102) следующим образом:

$$\begin{aligned} E_{\text{AV}} &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[E(y(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}E(y(\mathbf{x}))\right] \\ &\geq \mathbb{E}_{\mathbf{x}}\left[E\left(\sum_{m=1}^{M}\frac{1}{M}y(\mathbf{x})\right)\right] \\ &= E_{\text{COM}}, \end{aligned}$$

где $\lambda_i = 1/M$ для $i = 1, \ldots, M$ в (2.102), и получаем неявно определенные версии E_{AV} и E_{COM}, соответствующие $E(y(\mathbf{x}))$.

9.17. Чтобы доказать, что (9.67) является достаточным условием для (9.66), необходимо показать, что (9.66) следует из (9.67). Для этого рассмотрим фиксированный набор $y_m(\mathbf{x})$ и представим себе изменение α_m по всем возможным значениям, допускаемым (9.67), и рассмотрим значения, принимаемые $y_{\text{COM}}(\mathbf{x})$ в результате. Максимальное значение $y_{\text{COM}}(\mathbf{x})$ возникает, когда $\alpha_k = 1$, где $y_k(\mathbf{x}) \geq y_m(\mathbf{x})$ при $m \neq k$, и, следовательно, все $\alpha_m = 0$ при $m \neq k$. Аналогичный результат справедлив для минимального значения. Для других значений $\boldsymbol{\alpha}$

$$y_{\min}(\mathbf{x}) < y_{\text{COM}}(\mathbf{x}) < y_{\max}(\mathbf{x}),$$

поскольку $y_{\text{COM}}(\mathbf{x})$ является выпуклой комбинацией точек $y_m(\mathbf{x})$ таким образом, что

$$\forall m : y_{\min}(\mathbf{x}) \leq y_m(\mathbf{x}) \leq y_{\max}(\mathbf{x}).$$

Таким образом, (9.67) является достаточным условием для (9.66).

Доказательство того, что (9.67) является необходимым условием для (9.66), эквивалентно доказательству того, что (9.66) является достаточным условием для (9.67). Отсюда следует, что если (9.66) выполняется для любого выбора значений членов комитета $\{y_m(\mathbf{x})\}$, тогда будет выполняться (9.67). Предположим, без потери общей значимости, что α_k является наименьшим из значений α, т. е. $\alpha_k \leq \alpha_m$ при $k \neq m$. Затем рассмотрим $y_k(\mathbf{x}) = 1$ вместе с $y_m(\mathbf{x}) = 0$ для всех $m \neq k$. Тогда $y_{\min}(\mathbf{x}) = 0$, а $y_{\text{COM}}(\mathbf{x}) = \alpha_k$ и, следовательно, из (9.66) получаем

$\alpha_k \geq 0$. Поскольку α_k является наименьшим из значений α, то из этого следует, что все коэффициенты должны удовлетворять условию $\alpha_k \geq 0$. Подобным образом рассмотрим случай, когда $y_m(\mathbf{x}) = 1$ для всех m. Тогда $y_{\min}(\mathbf{x}) = y_{\max}(\mathbf{x}) = 1$, а $y_{\text{COM}}(\mathbf{x}) = \sum_m \alpha_m$. Из (9.66) следует, что $\sum_m \alpha_m = 1$, как и требовалось.

9.18. Из (3.2) распределение Бернулли для элементов матрицы отсева можно записать в виде

$$\text{Bern}(R_{ni}|\rho) = \rho^{R_{ni}}(1-\rho)^{1-R_{ni}}. \tag{377}$$

Следовательно, получаем

$$\begin{aligned}\mathbb{E}[R_{ni}] &= \sum_{R_{ni}\in\{0,1\}} \text{Bern}(R_{ni}|\rho)R_{ni} \\ &= \rho.\end{aligned} \tag{378}$$

Два элемента R_{ni} и R_{nj} будут независимыми, если $j \neq i$. Таким образом, для $j \neq i$ получаем

$$\begin{aligned}\mathbb{E}[R_{ni}R_{nj}] &= \sum_{R_{ni}\in\{0,1\}}\sum_{R_{nj}\in\{0,1\}} \text{Bern}(R_{ni}|\rho)\text{Bern}(R_{nj}|\rho)R_{ni}R_{nj} \\ &= \left(\sum_{R_{ni}\in\{0,1\}} \text{Bern}(R_{ni}|\rho)R_{ni}\right)\left(\sum_{R_{nj}\in\{0,1\}} \text{Bern}(R_{nj}|\rho)R_{nj}\right) \\ &= \rho^2,\end{aligned}$$

тогда как если $j = i$, получаем

$$R_{ni}R_{nj} = R_{ni}^2 = R_{ni}$$

и, следовательно,

$$\mathbb{E}[R_{ni}R_{nj}] = \mathbb{E}[R_{ni}] = \rho.$$

Объединяя эти выражения, получаем

$$\mathbb{E}[R_{ni}R_{nj}] = \delta_{ij}\rho + (1-\delta_{ij})\rho^2. \tag{379}$$

Чтобы найти ожидаемое значение функции ошибки (9.69), сначала разложим квадрат, чтобы получить

$$E(\mathbf{W}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk}^2 - 2\sum_{i=1}^{D} w_{ki}R_{ni}x_{ni} + \left(\sum_{i=1}^{D} w_{ki}R_{ni}x_{ni}\right)\left(\sum_{j=1}^{D} w_{kj}R_{nj}x_{nj}\right)\right\}.$$

Затем вычисляем математическое ожидание ошибки и подставляем математические ожидания элементов матрицы отсева, используя (378) и (379), чтобы получить

$$\mathbb{E}[E(\mathbf{W})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk}^2 - 2\rho\sum_{i=1}^{D} w_{ki}x_{ni} + \rho^2\left(\sum_{i=1}^{D} w_{ki}x_{ni}\right)^2\right\} + (\rho - \rho^2)\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D} w_{ki}^2 x_{ki}^2 \tag{380}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk} - \rho\sum_{i=1}^{D} w_{ki}x_{ni}\right\}^2 + \rho(1-\rho)\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D} w_{ki}^2 x_{ki}^2. \tag{381}$$

Наконец, можно найти решение для весов, которые минимизируют эту ожидаемую функцию ошибки, путем приравнивания производных по w_{ki} к нулю. Для этого удобнее работать с выражением (380), дающим

$$\begin{aligned}\frac{\partial}{\partial w_{ki}}\mathbb{E}[E(\mathbf{W})] &= -2\rho\sum_{n=1}^{N} x_{ni} + 2\rho^2\sum_{j=1}^{D} w_{kj}\left(\sum_{n=1}^{N} x_{nj}x_{ni}\right) \\ &\quad + 2\rho(1-\rho)w_{ki}\left(\sum_{n=1}^{N} x_{ni}^2\right) \\ &= -2\rho\sum_{n=1}^{N} x_{ni} + 2\sum_{j=1}^{D} w_{kj}\left\{\rho^2\left(\sum_{n=1}^{N} x_{nj}x_{ni}\right) + \delta_{ji}\rho(1-\rho)\left(\sum_{n=1}^{N} x_{ni}^2\right)\right\}.\end{aligned}$$

Это можно записать в матричной форме как

$$0 = -\mathbf{B} + \mathbf{WM}, \tag{382}$$

где $\mathbf{W}$ имеет элементы w_{ij}, $\mathbf{B}$ является диагональной матрицей, а элементы $\mathbf{B}$ и $\mathbf{M}$ задаются следующим образом:

$$B_{ii} = -2\rho\sum_{n=1}^{N} x_{ni}, \tag{383}$$

$$M_{ii} = 2\left\{\rho^2\left(\sum_{n=1}^{N} x_{nj}x_{ni}\right) + \delta_{ji}\rho(1-\rho)\left(\sum_{n=1}^{N} x_{ni}^2\right)\right\}. \tag{384}$$

Следовательно, минимизирующие веса задаются следующим образом:

$$\mathbf{W}^\star = \mathbf{BM}^{-1}. \tag{385}$$

Глава 10. Сверточные сети

10.1. Можно наложить ограничение $\|\mathbf{x}\|^2 = K$ с помощью множителя Лагранжа λ и максимизировать

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + \lambda(\|\mathbf{x}\|^2 - K). \tag{386}$$

Взяв градиент по $\mathbf{x}$ и приравняв его к нулю, получаем

$$\mathbf{w} + 2\lambda\mathbf{x} = 0, \tag{387}$$

что доказывает, что $\mathbf{x} = \alpha\mathbf{w}$, где $\alpha = -1/(2\lambda)$.

10.2. Представим входной массив в виде вектора $\mathbf{x} = (x_1, x_2 \ldots, x_5)^T$. Начнем со случая, когда имеется только один сверточный фильтр шириной 3, веса которого обозначим вектором $\mathbf{k} = (k_1; k_2; k_3)^T$. Если взглянуть на рис. 3, можно увидеть, что три выхода, которые представлены вектором $\mathbf{y} = (y_1; y_2; y_5)^T$, задаются следующим образом:

$$\mathbf{y} = \begin{bmatrix} x_1k_1 + x_2k_2 + x_3k_3 \\ x_2k_1 + x_3k_2 + x_4k_3 \\ x_3k_1 + x_4k_2 + x_5k_3 \end{bmatrix}. \tag{388}$$

Теперь необходимо найти такую матрицу $\mathbf{K}$, чтобы $\mathbf{y} = \mathbf{K}\mathbf{x}$. Очевидно, что это должна быть матрица 3×5, и каждый элемент K_{ij} задается вкладом x_j в y_i. Поэтому K задается следующим образом:

$$K = \begin{bmatrix} k_1 & k_2 & k_3 & 0 & 0 \\ 0 & k_1 & k_2 & k_3 & 0 \\ 0 & 0 & k_1 & k_2 & k_3 \end{bmatrix}. \tag{389}$$

Это пример матрицы Теплица (Toeplitz matrix), в которой каждая нисходящая диагональ слева направо является постоянной. Операции свертки в 1D всегда можно представить в виде умножения входного массива на матрицу Теплица.

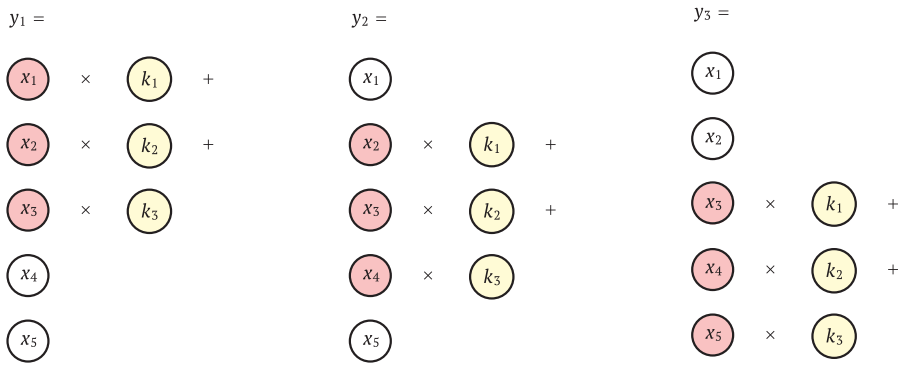

РИС. 3. *Изображение операции свертки фильтра* $\mathbf{k}$ *над входным массивом* $\mathbf{x}$ *с выходом* $\mathbf{y}$

10.3. Простое умножение матриц показывает, что свертка задается формулой

32	14	–18
–22	–24	12
22	6	18

10.4. Существует множество возможностей индексирования элементов I, K и C. Один из вариантов заключается в простом выборе индексов в $K(l;m)$ с диапазонами $1 \le l \le L$ и $1 \le m \le M$, что дает

$$C(j,k) = \sum_{l=1}^{L}\sum_{m=1}^{M} I(j+l, k+m)K(l,m). \tag{390}$$

Если подобным образом выбрать два индекса $\mathrm{I}(\cdot\ ;\ \cdot)$ для прогона от 1, ..., J и 1, ..., K соответственно, то получается, что $0 \le j \le J - L$ и $0 \le k \le K - M$. Для сверточной формы аналогично имеется

$$C(j,k) = \sum_{l=1}^{L}\sum_{m=1}^{M} I(j-l, k-m)K(l,m), \tag{391}$$

где теперь $\mathrm{L} + 1 \le j \le J + 1$ и $M + 1 \le k \le K + 1$, что легко проверить. Наконец, можно определить $\lambda = L - l + 1$ и $\mu = M - m + 1$, что позволяет переписать (391) в виде

$$C(j,k) = \sum_{\lambda=1}^{L}\sum_{\mu=1}^{M} \tilde{I}(j+\lambda, k+\mu)\tilde{K}(\lambda,\mu), \tag{392}$$

где определено

$$\tilde{I}(j+\lambda, k+\mu) = I(j-L+\lambda-1, k-M+\mu-1), \tag{393}$$

$$\tilde{K}(\lambda,\mu) = K(L-\lambda+1, M-\mu+1). \tag{394}$$

Представления свертки и кросс-корреляции различаются тем, что индексные переменные l и m, которые маркируют элементы ядра, идут от низких к высоким значениям или наоборот. В приложениях машинного обучения обычно не имеет значения, какая именно из этих форм будет использоваться, поскольку алгоритм будет учиться одному и тому же значению ядра в соответствующих местах, так что обучение с инвертированным представлением приведет к тому же ядру, но в обратном порядке значений.

10.5. Если подставить $z = x - y$ в (10.21), получим

$$F(x) = \int_{\infty}^{-\infty} G(x-z)k(z)\mathrm{d}z. \tag{395}$$

Если теперь провести дискретизацию переменных x и z в ячейки шириной Δ, можно приблизительно вычислить этот интеграл с помощью формулы

$$F(j\Delta) \simeq \sum_{l=-\infty}^{\infty} G(j\Delta - l\Delta)k(l\Delta)\Delta. \tag{396}$$

Теперь это можно записать в виде одномерной версии сверточного слоя, определенного формулой (10.19), в виде

$$C(j) \simeq \sum_{l=1}^{L} I(j-l)K(l), \tag{397}$$

где определены $C(j) = F(j\Delta)$, $I(l) = G(l\Delta)$ и $K(l) = k(l\Delta)\Delta$.

10.6. В разделе 10.2.3 было рассмотрено, что свертка изображения размерами $J \times K$ и с дополнительным заполнением P с фильтром размером $M \times M$ дает карту признаков размера $(J + 2P - M + 1) \times (K + 2P - M + 1)$. В данном случае, если подставить $P = (M - 1)/2$, получится, что размеры карты признаков задаются формулой $(J + (2((M - 1)/2) - M + 1) \times (K + (2((M - 1)/2) - M + 1)$, что в итоге упрощается до $J \times K$.

10.7. В случае без заполнения и со страйдом 1 ядро $M \times M$ будет свертываться $J - M + 1$ раз по горизонтали и $K - M + 1$ раз по вертикали, что дает $(J - M + 1) \times (K - M + 1)$ признаков. После применения заполнения P к каждому краю изображения получается новое изображение размером $(J + 2P) \times (K + 2P)$, и, следовательно, размерность слоя признаков будет $(J + 2P - M + 1) \times (K + 2P - M + 1)$. При применении страйда количество сверток в каждом измерении делится на страйд и с помощью оператора floor учитывается случай, когда в заданном направлении остается часть изображения, меньшая, чем страйд. Начальное 1 не делится на страйд, поскольку оно представляет первую операцию и, следовательно, не зависит от страйда. Это дает нам

$$\left\lfloor \frac{J + 2P - M}{S} + 1 \right\rfloor \times \left\lfloor \frac{K + 2P - M}{S} + 1 \right\rfloor \tag{398}$$

признаков.

10.8. Для упрощения предположим, что связи между заполняющими входами и признаками учитываются как связи. Кроме того, не учитываются связи максимального пула или функции активации. Поскольку каждый сверточный слой в сети VGG-16 использует фильтр 3×3, каждый узел в сверточном слое принимает количество входов, равное 9-кратному количеству каналов в предыдущем слое плюс 1 для смещения. Таким образом, первый сверточный слой имеет $224 \times 224 \times (3 \times 9 + 1) \times 64 = 96\ 337\ 920$ связей с предыдущим слоем. Количество связей для полносвязного слоя равно количеству входных признаков плюс 1 для смещения, умноженному на количество узлов, которое для первого полносвязного слоя равно $(7 \times 7 \times 512 + 1) \times 4096 = 102\ 764\ 544$. Количество связей для остальных слоев показано в табл. 1.

ТАБЛИЦА 1 *Количество соединений и обучаемых параметров в каждом слое сети VGG-16*

Слой	Связи	Обучаемых параметров
Сверточный 1	96 337 920	1792
Сверточный 2	1 852 899 328	36 928
Сверточный 3	926 449 664	73 856
Сверточный 4	1 852 899 328	147 584
Сверточный 5	926 449 664	295 168
Сверточный 6	1 852 899 328	590 080
Сверточный 7	1 852 899 328	590 080
Сверточный 8	926 449 664	1 180 160
Сверточный 9	1 852 899 328	2 359 808
Сверточный 10	1 852 899 328	2 359 808
Сверточный 11	462 522 368	2 359 808
Сверточный 12	462 522 368	2 359 808
Сверточный 13	462 522 368	2 359 808
Полносвязный 1	102 764 544	102 764 544
Полносвязный 2	16 781 312	16 781 312
Полносвязный 3	4 097 000	4 097 000
Итого	15 504 292 840	138 357 544

Количество обучаемых параметров в сверточном слое не зависит от высоты и ширины этого слоя. Для слоя с фильтром 3×3 количество параметров для данного ядра равно 9, умноженному на количество каналов в предыдущем слое, плюс 1 для смещения. Для данного слоя количество таких ядер равно количеству каналов. Так, например, первый сверточный слой имеет $64 \times (3 \times 3 \times 3 + 1) = 1792$ обучаемых параметра. Для полносвязного слоя количество параметров равно количеству связей, поскольку нет общих весов. Количество обучаемых параметров для каждого слоя также показано в табл. 1.

10.9. Операция свертки может быть записана в виде

$$C(j,k) = \sum_{l=1}^{L} \sum_{m=1}^{M} I(j-l, k-m) K(l,m), \tag{399}$$

где $j = 1, \ldots, J$ и $k = 1, \ldots, K$. Ядро K проходит по всему изображению I, образуя общее количество позиций $(J - L + 1) \times (K - M + 1)$, и для каждой позиции количество операций равно $L \times M$. Таким образом, общее количество операций равно

$$(J - L + 1)(K - M + 1)LM. \tag{400}$$

Теперь предположим, что ядро является разделимым, т. е. факторизуется в виде

$$K(l,m) = F(l)G(m). \tag{401}$$

Подставляя (401) в (399), получаем

$$C(j,k) = \sum_{l=1}^{L} F(l) \sum_{m=1}^{M} I(j-l, k-m) G(m). \tag{402}$$

Для начала рассмотрим суммирование по m. Это включает одномерное ядро $G(m)$, которое должно быть просканировано по изображению для общего числа $J \times (K - M + 1)$ позиций, и в каждой позиции необходимо выполнить M операций, что дает общее число операций, равное $(K - M + 1)JM$. Это приводит к промежуточному массиву размерности $J \times (K - M + 1)$. Теперь выполняется суммирование по l, которое также является сверткой с одномерным ядром $F(l)$. Число позиций для этого ядра задается формулой $(K - M + 1) \times (J - L + 1)$, и в каждой позиции необходимо выполнить L операций, что дает общее число $(K - M + 1)(J - L + 1)L$. Таким образом, общее число операций задается выражением

$$(K - M + 1)JM + (K - M + 1)(J - L + 1)L. \tag{403}$$

Чтобы убедиться, что это позволяет сократить вычисления, рассмотрим случай, когда изображение большое по сравнению с размером ядра, так что $J \gg L$ и $K \gg M$. Тогда уравнение (400) приблизительно задается $JKLM$, а уравнение (403) задается $JK(L + M)$. Здесь также следует отметить, что, помимо экономии вычислительных ресурсов, разделяемое ядро требует меньше памяти. Тем не менее, поскольку оно ограничивает форму ядра, это может привести к значительному снижению точности обобщения.

10.10. Производные функции затрат по отношению к значению активации могут быть оценены с помощью обратного распространения, что соответствует применению правила цепочки в математическом анализе. Обратное распространение начинается с производных функции затрат по отношению к локальным активациям, которые для функции затрат, определенной формулой (10.12), задаются формулой

$$\delta_{ijk} = \frac{\partial F(\mathbf{I})}{\partial a_{ijk}} = 2a_{ijk} \tag{404}$$

и, следовательно, с точностью до коэффициента 2 задаются самими значениями активации. Затем эти значения обратно распространяются по сети с применением (8.13) до тех пор, пока не будет достигнут входной слой. Этот входной слой представляет изображение, а связанные с ним значения δ соответствуют производным функции затрат (10.12) по отношению к значениям пикселей, как и требовалось.

10.11. Рассмотрим схему «одноразового горячего» (1-hot) кодирования для классов объектов C с использованием бинарных переменных $y_i \in \{0, 1\}$, где $i = 1, \ldots, C$ и где $y_i = 1$ означает наличие объекта из класса i. Затем введем дополнительный класс с бинарной переменной $y_{C+1} \in \{0, 1\}$, где $y_{C+1} = 1$ означает, что

в изображении нет объектов ни из одного из заданных классов. Поскольку предполагается, что данное изображение содержит либо объект из одного из классов, либо не содержит объектов, переменные $y_1, \ldots, y_{C+1}$ формируют кодировку 1-hot, в которой все переменные имеют значение 0, за исключением одной переменной, принимающей значение 1. Можно обучить модель $f(\cdot)$ принимать изображение в качестве входных данных и возвращать вероятностное распределение по этим переменным. Сумма этих вероятностей должна равняться единице:

$$\sum_{i=1}^{C+1} p_f(y_i = 1) = 1. \tag{405}$$

Теперь вместо этого предположим, что вводим бинарную переменную $b \in \{0, 1\}$, такую что $b = 1$ означает, что объект (любого класса) присутствует на изображении, а $b = 0$ означает, что объект отсутствует. Можно обучить модель $g(\cdot)$ выводить вероятностное распределение по этой переменной, так что

$$p_g(b = 1) + p_g(b = 0) = 1. \tag{406}$$

Также введем бинарные переменные $z_1, \ldots, z_C \in \{0, 1\}$ для прогнозирования класса объекта при условии, что объект присутствует. Эти переменные имеют кодировку 1-hot. Затем обучим связанную модель $h(\cdot)$ для вывода вероятностного распределения, удовлетворяющего

$$\sum_{i=1}^{C} p_h(z_i = 1) = 1. \tag{407}$$

Чтобы связать эти наборы вероятностей, мы можем использовать правило произведения в виде

$$\begin{aligned} &p(\text{объект присутствует и класс } i) \\ &\quad = p(\text{класс } i | \text{объект присутствует}) p(\text{объект присутствует}), \end{aligned} \tag{408}$$

что дает следующие результаты:

$$p_f(y_i = 1) = p_h(z_i = 1) p_g(b = 1) \quad i = 1, \ldots, C, \tag{409}$$

$$p_f(y_{C+1} = 1) = p_g(b = 0). \tag{410}$$

10.12. Сначала отметим, что для вычисления скалярного произведения двух N-мерных векторов требуется N умножений и $N - 1$ сложений, что в сумме дает $2N - 1$ вычислительных шагов.

Для сети, представленной на рис. 10.22, количество вычислительных шагов, необходимых для вычисления первой операции свертки, определяется количеством признаков во втором слое, которое равно $4 \times 4 = 16$, умноженным на количество шагов, необходимых для вычисления выхода фильтра. Поскольку размер фильтра равен $3 \times 3 = 9$, каждое вычисление фильтра тре-

бует 9 × 8 = 72 шага. Таким образом, общее количество вычислительных шагов для первого слоя свертки составляет 16 × 72 = 1152. Для одной оценки фильтра максимального пулинга 2 × 2 требуется 3 вычислительных шага. Это умножается на количество таких операций, необходимых для слоя максимального пулинга, которое равно 4, что дает в общей сложности 12 шагов. Полносвязный слой эквивалентен скалярному произведению двух векторов размерности 4 и, следовательно, требует 4 + 3 = 7 операций. Таким образом, для одной оценки этой сети требуется в общей сложности 1152 + 12 + 7 = 1171 вычислительных шагов.

Для сети на рис. 10.23 во втором слое имеется 6 × 6 = 36 признаков, каждый из которых требует 9 × 8 = 72 вычислений в слое свертки, что дает в общей сложности 36 × 72 = 2592 вычислительных шага. Слой максимального пулинга имеет 9 узлов и, следовательно, требует 9 × 3 = 27 вычислений. Наконец, полносвязный слой требует 4 × 7 = 28 вычислений, что дает в общей сложности 2592 + 27 + 28 = 2647 операций для одного прохода по всей сети.

Таким образом, повышение эффективности использования одного прохода по второй сети, по сравнению с 8 проходами по первой сети, равно 1171 × 8/2647 = 3,54.

10.13. Вектор входных данных с заполнением задается следующим образом:

$$\mathbf{x} = \begin{pmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ 0 \end{pmatrix}. \tag{411}$$

Для фильтра с элементами (w_1, w_2; w_3) и страйдом 2 выходной вектор будет двумерным и может быть записан как $y = (y_1, y_2)^T$, где элементы задаются следующим образом:

$$y_1 = w_2 x_1 + w_3 x_2, \tag{412}$$

$$y_2 = w_1 x_2 + w_2 x_3 + w_3 x_4. \tag{413}$$

Эту операцию свертки можно записать с помощью матричной нотации в виде

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{414}$$

где матрица $\mathbf{A}$ задается следующим образом:

$$\mathbf{A} = \begin{pmatrix} w_1 & w_2 & w_3 & 0 & 0 & 0 \\ 0 & 0 & w_1 & w_2 & w_3 & 0 \end{pmatrix}. \tag{415}$$

Теперь рассмотрим операцию повышающей дискретизации с использованием фильтра (w_1, w_2, w_3), действующего на вектор $\mathbf{z} = (z_1, z_2)^{\mathrm{T}}$ со страйдом 2. Результирующий шестимерный вектор $\mathbf{h} = (h_1, h_2, h_3, h_4, h_5, h_6)$ имеет элементы, заданные следующим образом:

$$h_1 = w_1 z_1, \tag{416}$$

$$h_2 = w_2 z_1, \tag{417}$$

$$h_3 = w_3 z_1 + w_1 z_2, \tag{418}$$

$$h_4 = w_2 z_2, \tag{419}$$

$$h_5 = w_3 z_2, \tag{420}$$

$$h_6 = 0. \tag{421}$$

$$\tag{422}$$

Это можно записать с помощью матричной нотации в виде

$$\mathbf{h} = \mathbf{B}\mathbf{z}, \tag{423}$$

где матрица $\mathbf{B}$ задается следующим образом:

$$\mathbf{B} = \begin{pmatrix} w_1 & 0 \\ w_2 & 0 \\ w_3 & w_1 \\ 0 & w_2 \\ 0 & w_3 \\ 0 & 0 \end{pmatrix}. \tag{424}$$

При внимательном рассмотрении можно заметить, что $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$, и, следовательно, повышение частоты дискретизации можно рассматривать как пример транспонирующей свертки.