

Лекция 2

EDA и уменьшение размерности

Лектор

Елизавета Власова

Практики

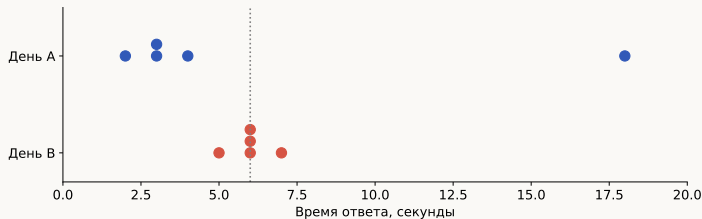
Кирилл Захаров

Камила Насибуллина

Михаил Подсытник

Игорь Толстокулаков

Среднее, медиана и разброс: пример



День А: 2, 3, 3, 4, 18 с. День В: 5, 6, 6, 6, 7 с.

Среднее в обоих случаях — 6 с. Как различаются медиана и разброс? Какое время лучше описывает типичный запрос в день А?

Среднее и медиана при правом хвосте

$$\text{median}(A) = 3 \text{ с}, \quad \text{median}(B) = 6 \text{ с}.$$

$$s_A^2 = 45.5 \text{ с}^2, \quad s_B^2 = 0.5 \text{ с}^2.$$

- ▶ В день A четыре из пяти ответов занимают 2–4 с; среднее сдвигает ответ в 18 с.
- ▶ Для типичного запроса здесь информативнее медиана; среднее нужно, например, для суммарного времени.
- ▶ Перед преобразованием смотрим на форму распределения, единицы и редкие значения.

Общий пайплайн анализа данных

1. **Вопрос и критерий ответа:** что хотим узнать и о какой совокупности?
2. **Данные и их качество:** сбор, единица наблюдения, пропуски, ошибки.
3. **EDA:** распределения, группы, связи, необычные наблюдения.
4. **Подготовка:** масштаб признаков и представление данных.
5. **Проверка:** гипотеза и неопределённость либо модель и валидация.
6. **Вывод:** ответ на вопрос, ограничения, следующий шаг.

Сегодня: масштаб → статистический вывод → корреляции → PCA → t-SNE.

Масштабирование данных

Единицы измерения и преобразования
признаков

Масштаб признаков и расстояния

Объект	Возраст, годы	Доход, тыс. руб.
A	20	50
B	30	51
C	21	60

$$d(A, B) = \sqrt{10^2 + 1^2} = d(A, C).$$

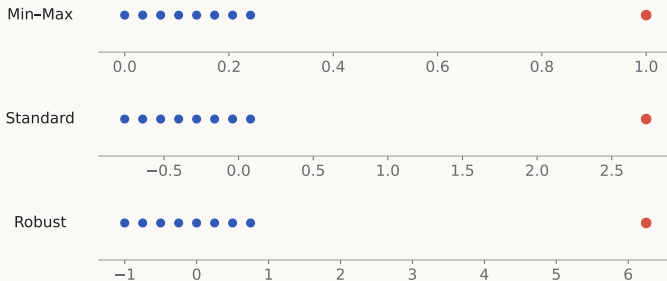
Если доход записать в рублях, вклад дохода станет в 10^6 раз больше в **квадрате** расстояния.

Методы масштабирования

Метод	Формула	Опора
Min-Max	$\frac{X - X_{\min}}{X_{\max} - X_{\min}}$	Минимум и максимум
Standard	$\frac{X - \mu}{\sigma}$	Среднее и масштаб
Robust	$\frac{X - \text{median}(X)}{Q_{.75} - Q_{.25}}$	Медиана и IQR

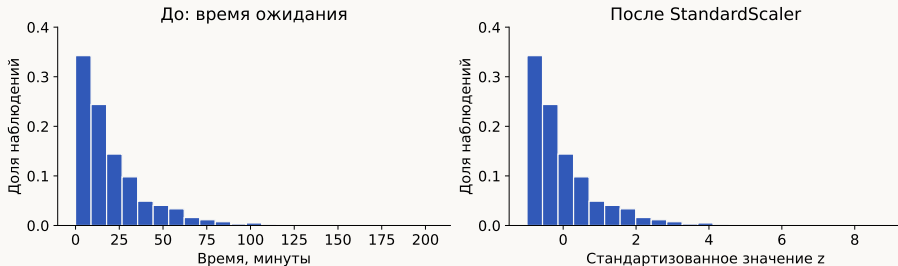
Параметры вычисляем по обучающей выборке, отдельно для каждого признака. Нулевой знаменатель требует отдельной обработки.

RobustScaler и выбросы



Одни и те же значения: 1, 2, ..., 8, 30. Красная точка остаётся крайней.

Стандартизация и форма распределения



Пример: 2000 значений из экспоненциального распределения (среднее 20 мин). После преобразования $\bar{z} = 0$, $\text{Var}(z) = 1$, но правый хвост остаётся.

Задача: Min-Max для нового объекта

На train: $x_{\min} = 10$, $x_{\max} = 30$.

- ▶ Во что Min-Max преобразует $x = 20$?
- ▶ А новый объект $x = 40$?
- ▶ Нужно ли пересчитать максимум по test?

Min-Max: разбор задачи

$$20 \mapsto \frac{20 - 10}{30 - 10} = 0.5, \quad 40 \mapsto \frac{40 - 10}{30 - 10} = 1.5.$$

- ▶ Значение за пределами $(0, 1)$ допустимо.
- ▶ Пересчёт параметров по `test` меняет преобразование и раскрывает информацию о `test`.
- ▶ Масштаб выбирают под метод: расстояния, РСА и регуляризация чувствительны к нему.

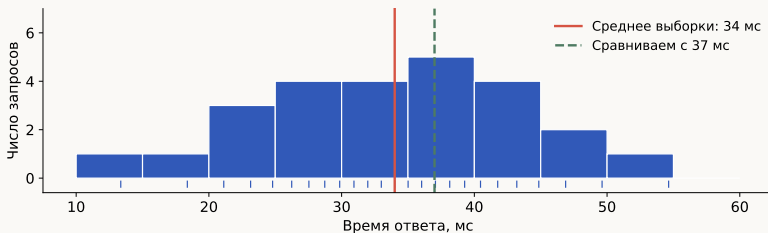
2

Статистический анализ

Распределения, доверительные интервалы и
проверка гипотез

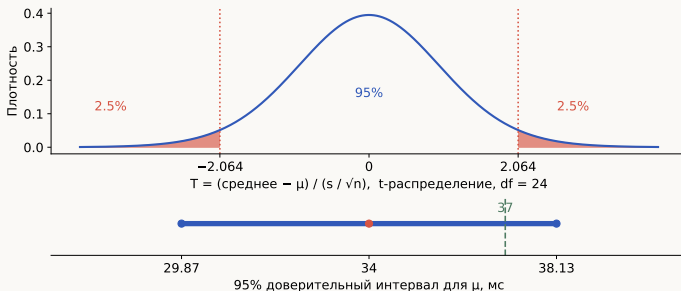
Выборка времени ответа

25 запросов: $\bar{X} = 34$ мс, выборочное СКО $s = 10$ мс.



Изменилось ли среднее время ответа по сравнению с 37 мс?

Доверительный интервал среднего



$$\bar{X} \pm t_{24,0.975} s / \sqrt{n} = 34 \pm 2.064 \cdot 2 = [29.87; 38.13] \text{ мс.}$$

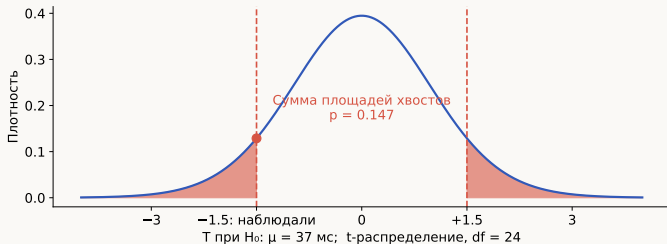
При независимых нормальных наблюдениях 95% интервалов, построенных этой процедурой, покрывают μ .

Статистические гипотезы

Вопрос	Нулевая гипотеза	Альтернатива
Есть изменение?	$H_0 : \mu = 37$	$H_1 : \mu \neq 37$
Стало медленнее?	$H_0 : \mu \leq 37$	$H_1 : \mu > 37$

- ▶ Статистика измеряет расхождение данных с нулевой моделью.
- ▶ Её распределение при H_0 переводит расхождение в вероятность.

p-value и хвосты распределения



$$T_{\text{obs}} = (34 - 37)/(10/\sqrt{25}) = -1.5.$$

$p = P_{H_0}(|T| \geq 1.5) \approx 0.147$: при $\alpha = 0.05$ гипотезу $\mu = 37$ не отвергаем.

Ошибки I и II рода

Решение	В мире H_0 верна	В мире H_0 неверна
Отвергнуть H_0	Ошибка I рода	Обнаружить эффект
Не отвергать H_0	Корректное решение	Ошибка II рода

α ограничивает вероятность ошибки I рода.

Мощность $1 - \beta$ — вероятность обнаружить заданную альтернативу.

Честный ли кубик?

После 60 независимых бросков получены частоты:

Грань	1	2	3	4	5	6
Частота	8	9	12	11	13	7

- ▶ Сформулируйте нулевую гипотезу.
- ▶ Какие частоты ожидаются при этой гипотезе?
- ▶ Как одним числом измерить расхождение наблюдений и ожиданий?

Распределение исходов честного кубика



60 независимых бросков; $H_0 : p_1 = \dots = p_6 = 1/6$. Ожидаемая частота каждой грани $E_i = 10$.

Критерий χ^2 : наблюдения и ожидания

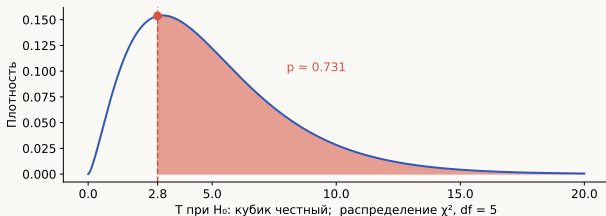
Грань	1	2	3	4	5	6
Наблюдали O_i	8	9	12	11	13	7
Ожидали E_i	10	10	10	10	10	10
$O_i - E_i$	-2	-1	2	1	3	-3
$(O_i - E_i)^2 / E_i$	0.4	0.1	0.4	0.1	0.9	0.9

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i} = 2.8.$$

Простая сумма отклонений равна нулю. Квадраты учитывают величину расхождения; деление на E_i — масштаб колебаний частот.

Распределение χ^2

Если Z_j независимы и $Z_j \sim N(0, 1)$, то $\sum_{j=1}^k Z_j^2 \sim \chi_k^2$.



У частот кубика $\sum O_i = 60$: остаётся $6 - 1 = 5$ степеней свободы. При H_0 статистика T приближённо имеет распределение χ_5^2 .

$p = P_{H_0}(T \geq 2.8) \approx 0.731$: гипотезу о честном кубике не отвергаем.

Задача: интерпретация p-value

В другом эксперименте получили $p = 0.20$. Какой вывод корректен?

- ▶ А. Вероятность верности H_0 равна 20%.
- ▶ В. При $\alpha = 0.05$ данных недостаточно, чтобы отвергнуть H_0 .
- ▶ С. Эффекта точно нет.

Интерпретация p-value: разбор

Верный ответ: В.

- ▶ p-value не даёт вероятность гипотезы.
- ▶ В отчёте нужны величина эффекта, интервал неопределённости и условия измерения.
- ▶ Если искать много связей и показывать только значимые, ложные находки накапливаются.

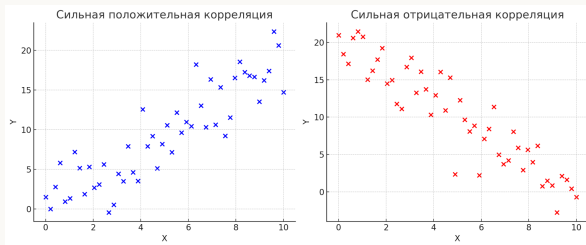
3

Корреляции

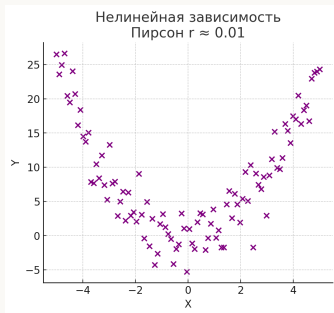
Линейные и ранговые связи между признаками

Линейная корреляция Пирсона

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$



Нелинейная зависимость и корреляция

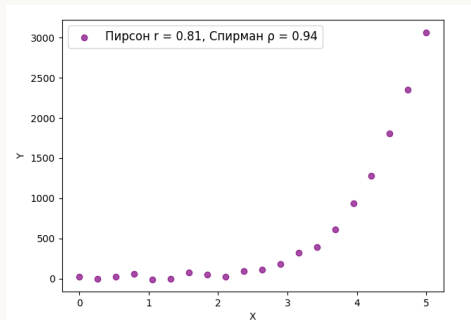


Зависимость есть.

Прямая её не описывает.

Перед выводом по коэффициенту
смотрим на форму связи.

Ранговая корреляция Спирмена



$$\rho_s = \text{corr}(\text{rank}(x), \text{rank}(y)).$$

Монотонная связь может быть нелинейной.

При совпадениях используем средние ранги.

Пример расчёта корреляции Спирмена

Наблюдение	1	2	3	4	5	6	7	8
Время	12	15	17	18	20	21	22	26
Возраст	14	25	20	35	45	30	60	95
R_x	1	2	3	4	5	6	7	8
R_y	1	3	2	5	6	4	7	8

$$\sum_i (R_{x,i} - R_{y,i})^2 = 8, \quad \rho_s = 1 - \frac{6 \cdot 8}{8(8^2 - 1)} \approx 0.905.$$

Последняя формула применима без совпадающих рангов.

Задача: корреляция после преобразования

- ▶ Время из предыдущей таблицы перевели из минут в секунды. Что станет с r и ρ_s ?
- ▶ А если заменить время его квадратом: какие ранги изменятся?

Корреляции и преобразования признаков

Преобразование времени	Pearson r	Spearman ρ_s
$x \mapsto 60x$	Не изменится	Не изменится
$x \mapsto x^2$, здесь $x > 0$	Может измениться	Не изменится

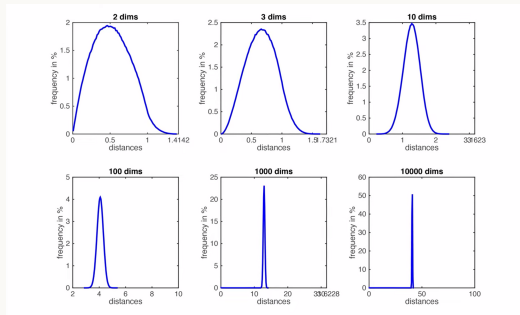
- ▶ Положительное линейное масштабирование сохраняет оба коэффициента.
- ▶ Строго возрастающее преобразование сохраняет ранги, но может изменить линейную форму.

4

Снижение размерности и PCA

Зачем уменьшать число признаков и как
выбрать новые оси

Проклятие размерности



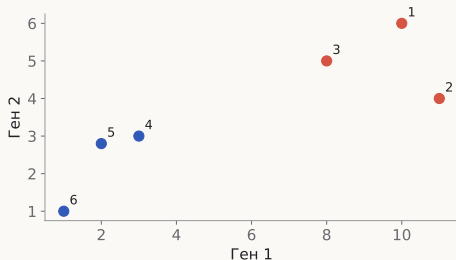
На графиках — распределения попарных расстояний при разном числе признаков.

Относительный разброс расстояний уменьшается.

«Близкий» и «далёкий» соседи становятся менее различимы.

Эффект зависит от распределения данных и метрики.

Два связанных признака и одна координата



Шесть образцов, два уровня экспрессии.

Если оба признака меняются вместе, можно описывать положение одной координатой.

Какую информацию потеряем?

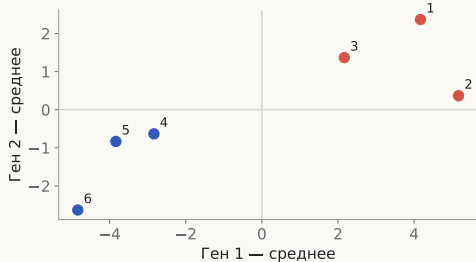
Данные для примера PCA

Образец	1	2	3	4	5	6
Ген 1	10	11	8	3	2	1
Ген 2	6	4	5	3	2.8	1

$$\mu = (5.833, 3.633), \quad X_c = X - \mathbf{1}\mu^\top.$$

Здесь X — матрица 6×2 : строки — объекты, столбцы — признаки. В примере используем центрирование без масштабирования.

Центрирование данных



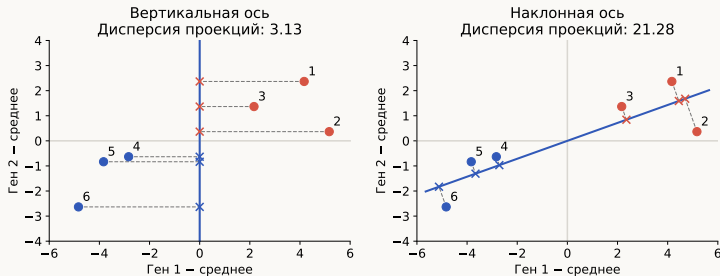
Для образца 1:

$$x_1 = (10, 6)$$

$$x_{c,1} = (4.167, 2.367).$$

Среднее каждого столбца X_c равно нулю.

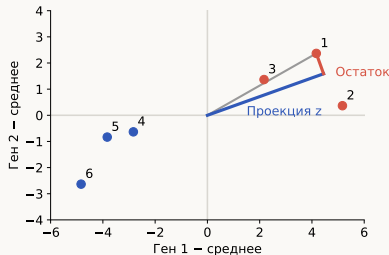
Сравнение направлений проекции



Крестики — ортогональные проекции. Для каждой оси измеряем разброс координат проекций.

Источник: StatQuest, PCA Step-by-Step

Дисперсия проекций и ошибка восстановления



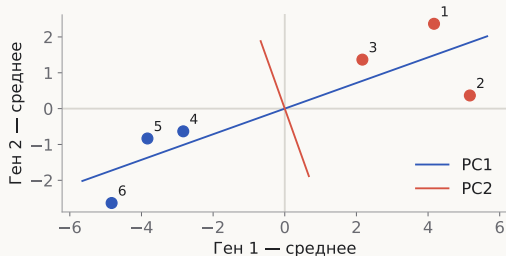
$$\|x_i\|^2 = z_i^2 + \|r_i\|^2.$$

Здесь x_i центрированы, r_i — остаток.
Сумма $\sum_i \|x_i\|^2$ фиксирована.

Максимизируем $\sum_i z_i^2$ —
минимизируем $\sum_i \|r_i\|^2$.

Это квадраты **перпендикулярных**
расстояний до оси.

Первая и вторая главные компоненты

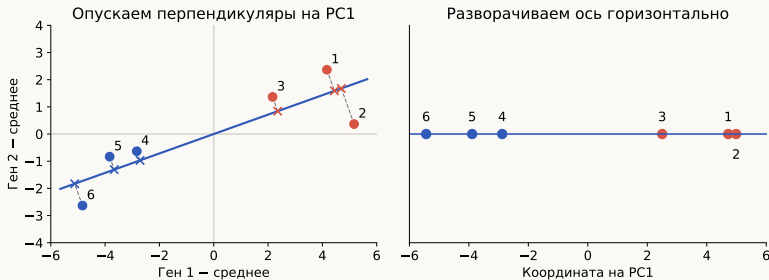


PC1 — наибольшая дисперсия.

PC2 — перпендикулярна PC1.

В двух измерениях PC2 сохраняет оставшуюся дисперсию.

Проекция и новые координаты



Номера и цвета обозначают те же образцы. Координата — **знаковое** расстояние от центра вдоль PC1.

Источник: StatQuest, PCA Step-by-Step

Линейная комбинация признаков

$$Z_i = w_1(X_{i1} - \mu_1) + w_2(X_{i2} - \mu_2), \quad \|w\| = 1.$$

Было	Стало
Координаты по исходным признакам	Координаты по новым направлениям
Два числа на объект	Одно число при $k = 1$

PCA выбирает направления с максимальной дисперсией проекций.

Ковариационная матрица

Для d признаков есть d^2 попарных ковариаций — таблица $d \times d$.

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \mu_j)(x_{ik} - \mu_k).$$

$$C = \begin{pmatrix} s_1^2 & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & s_2^2 \end{pmatrix} \approx \begin{pmatrix} 18.967 & 6.487 \\ 6.487 & 3.127 \end{pmatrix}.$$

Строка и столбец обозначают признаки; на диагонали — их дисперсии.

Для центрированных данных $C = X_C^T X_C / (n-1)$; дисперсия проекции на w равна $w^T C w$.

Собственные числа и векторы

$$Cv = \lambda v, \quad v \neq 0.$$

- ▶ v — **собственный вектор**: действие C оставляет его на той же прямой.
- ▶ λ — **собственное число**: во сколько раз C растягивает этот вектор.
- ▶ Для ковариационной матрицы C : $\lambda \geq 0$; собственные векторы можно выбрать ортонормированными.

Связь с PCA при $\|v\| = 1$:

$$z = X_C v, \quad s_z^2 = v^T C v = \lambda.$$

Почему в PCA возникают собственные векторы

Ищем единичную ось с максимальной дисперсией проекций:

$$\max_w w^T C w, \quad w^T w = 1.$$

Ограничение на длину учитываем множителем λ :

$$\mathcal{L}(w, \lambda) = w^T C w - \lambda(w^T w - 1).$$

$$\nabla_w \mathcal{L} = 2Cw - 2\lambda w = 0 \implies Cw = \lambda w.$$

Получили уравнение собственного вектора. В этой точке дисперсия $w^T C w = \lambda$; максимум даёт наибольшее собственное число.

Главные компоненты и собственные числа

$$\max_{\|w\|=1} w^T C w = \lambda_1, \quad \lambda_1 \geq \lambda_2 \geq 0.$$

Компонента	Направление	Дисперсия
PC1	(0.9417, 0.3364)	$\lambda_1 \approx 21.284$
PC2	(-0.3364, 0.9417)	$\lambda_2 \approx 0.809$

PC1 — собственный вектор для наибольшего λ ; PC2 — следующая ортогональная ось. $Z = X_C V_k$ — координаты на выбранных осях.

Пример вычисления координаты PC1

Для образца 1:

$$z_{11} = x_{c,1}^T v_1 \approx 4.167 \cdot 0.9417 + 2.367 \cdot 0.3364 \approx 4.720.$$



$Z = X_c V_k$ имеет размер $n \times k$: столько же объектов, меньше координат.

Задача: проекция нового объекта

Используйте $\mu = (5.833, 3.633)$ и $v_1 = (0.9417, 0.3364)$.

- ▶ Новый объект $x = (7, 4)$. Сначала найдите x_c .
- ▶ Затем вычислите $z = x_c^\top v_1$.
- ▶ Надо ли обучать PCA заново для этого объекта?

Проекция нового объекта: разбор

$$x_c = (1.167, 0.367), \quad z \approx 1.167 \cdot 0.9417 + 0.367 \cdot 0.3364 \approx 1.222.$$

- ▶ Берём среднее и направления, оценённые на train.
- ▶ Повторный fit создаст другую систему координат.
- ▶ Все последующие объекты преобразуем одним обученным отображением.

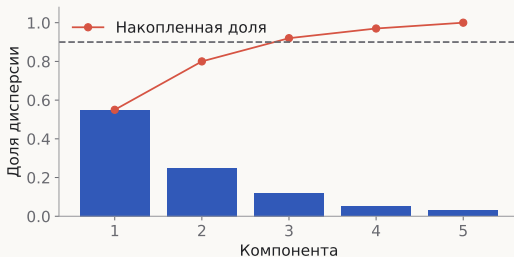
Объяснённая дисперсия

$$\text{EVR}_j = \frac{\lambda_j}{\sum_{\ell=1}^d \lambda_{\ell}}.$$

Наш пример	Сохранённая дисперсия	Потерянная дисперсия
Только PC1	96.34%	3.66%
PC1 и PC2	100%	0%

Это доля дисперсии выбранного представления X , а не доля сохранённой информации о целевой переменной.

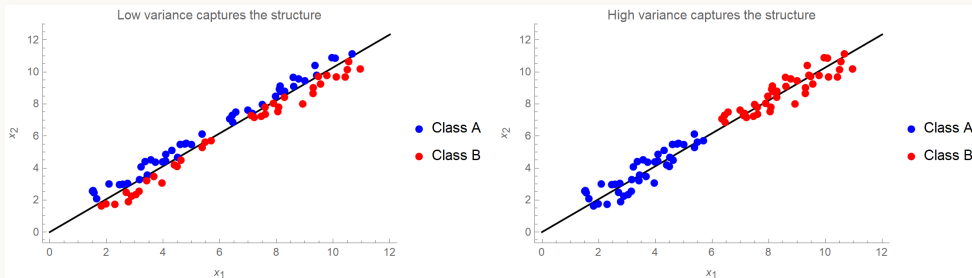
Число главных компонент



Условный пример: для порога 90% достаточно $k = 3$ (92%).

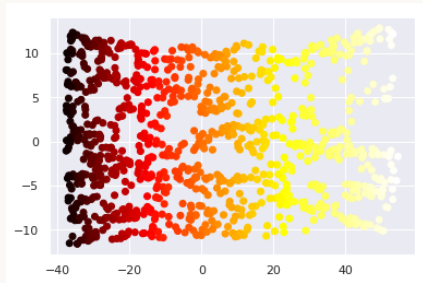
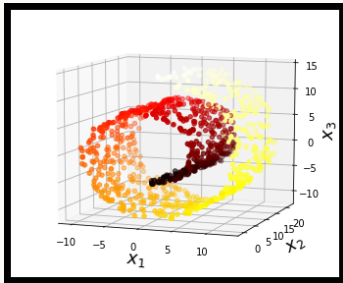
Для визуализации берём 2–3 оси; для модели проверяем качество на validation.

Дисперсия и разделение классов



Направление полезное для разделения классов может иметь малую дисперсию. PCA не использует y .

Ограничения линейной проекции



Линейная проекция может наложить разные участки свёрнутой поверхности.

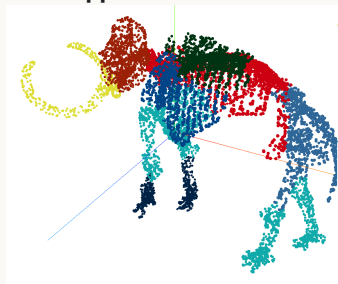
Подготовка данных перед PCA

Шаг	Зачем он нужен
Центрирование	PCA описывает отклонения от среднего
Масштабирование	Разные единицы меняют ковариации и оси
Проверка выбросов	Квадраты отклонений усиливают их влияние
Fit на train	Новые объекты используют те же параметры

Для общего осмысленного масштаба центрирования может быть достаточно. При разных единицах обычно рассматривают стандартизацию.

Мамонт: исходные данные и t-SNE

Исходное облако в 3D

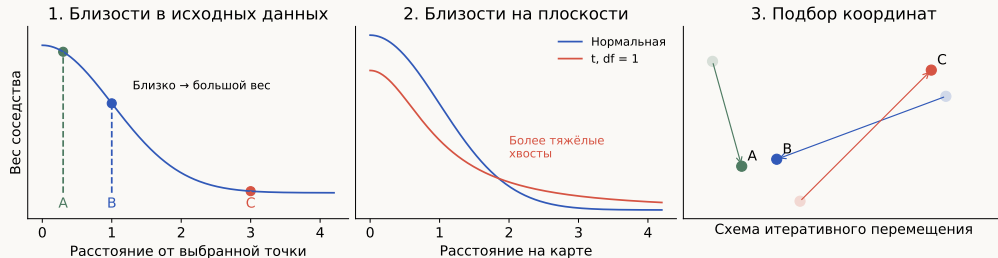


Проекция t-SNE в 2D



Цвета позволяют сопоставить участки облака. Локальные группы различимы, но форма мамонта и глобальные расстояния искажены. Параметр исходного примера: perplexity = 500.

t-SNE: основные шаги



1. Из расстояний получаем вероятности соседства p_{ij} .
2. На карте считаем близости q_{ij} с тяжёлыми хвостами t-распределения.
3. Перемещаем точки, уменьшая расхождение P и Q . Это не сохранение всех расстояний.

Источник: StatQuest: t-SNE, Clearly Explained

Интерпретация карты t-SNE



t-SNE адаптирует локальный масштаб. Площади групп и расстояния между ними нельзя напрямую трактовать как исходную геометрию.

Для проверки: другие seed и perplexity, исходные признаки и технические партии данных.

PCA и t-SNE: задачи и ограничения

Цель	Начальная точка	Что проверяем
Сжать линейную избыточность	PCA	EVR и качество задачи
Посмотреть локальные соседства	t-SNE	Устойчивость карты
Предсказывать y	Сравнить с исходными X	Метрику на отложенных данных

Два измерения — ограничение рисунка, а не обещание, что в данных только два существенных фактора.

Проверка понимания

- ▶ «После StandardScaler данные стали нормальными».
- ▶ «PCA сохранил 95% дисперсии — значит, 95% сигнала о y ».
- ▶ «Два острова t-SNE доказывают два класса».
- ▶ «При $p = 0.2$ нулевая гипотеза доказана».

Разбор итоговых вопросов

Про что утверждение	Корректная граница
Масштаб	Не меняет форму распределения
Дисперсия X	Не измеряет полезность для y
Геометрия карты	Не подтверждает классы
p-value	Не даёт вероятность гипотезы

Теоретический минимум

- ▶ Масштабирование: Min–Max, Standard, Robust; параметры по train.
- ▶ Статистический вывод: ДИ, H_0/H_1 , p-value, ошибки I/II рода; χ^2 как мера расхождения частот.
- ▶ Pearson и Spearman: линейная и монотонная связь, ограничения.
- ▶ Проклятие размерности: концентрация расстояний; роль избыточных признаков.
- ▶ PCA: $C, C v = \lambda v$; v — ось, λ — дисперсия. Центрирование, проекция, EVR и выбор k .
- ▶ t-SNE: локальные соседства и ограничения интерпретации карты.

Дальше

Ко(т)нец лекции



На следующей лекции

Линейные модели

Функции потерь, обучение и
регуляризация.

Выбор статистического критерия

Данные и вопрос	Возможный критерий
Одна выборка: среднее против μ_0	Одновыборочный t
Две независимые: средние	Welch t
Парные измерения: знак изменения	Парный критерий знаков
Независимые: распределения	Манна–Уитни
Частоты по заданным категориям	χ^2 согласия

Выбор зависит от независимости, типа вопроса и предпосылок, а не только от теста на нормальность.

Одновыборочный t-тест и тест Уэлча

Одна выборка

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1.$$

Точное t-распределение: iid нормальные наблюдения.

s_x^2, s_y^2 — выборочные дисперсии с знаменателями $n - 1, m - 1$. Парные данные анализируют через разности, а не как независимые.

Две независимые: Welch

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_x^2/n + s_y^2/m}}.$$

Не требует равенства дисперсий; степени свободы приближённые.

Критерий знаков

$$D_i = X_i - Y_i, \quad S_+ = \sum_{i=1}^n \mathbf{1}[D_i > 0].$$

- ▶ Для непрерывных независимых разностей при H_0 с медианой 0:
 $S_+ \sim \text{Bin}(n, 1/2)$.
- ▶ Нулевые разности исключаем; n — число ненулевых разностей.
- ▶ Пример: 7 положительных и 8 отрицательных; двустороннее $p = 1$.
- ▶ Для одной выборки вместо Y_i используем заданную медиану m_0 .

Критерий Манна–Уитни

Размеры выборок: n и m . Присваиваем общие ранги. R_x — сумма рангов первой выборки.

$$U_x = R_x - \frac{n(n+1)}{2}, \quad U_y = nm - U_x.$$

- ▶ При H_0 распределения одинаковы; независимость наблюдений обязательна.
- ▶ Это не универсальный тест равенства средних или медиан.
- ▶ Интерпретация как сдвиг положения требует дополнительных условий о форме распределений.

Материалы и проверка формул

- ▶ StatQuest: PCA, Step-by-Step. Геометрическая последовательность PCA.
- ▶ StatQuest: t-SNE, Clearly Explained.
- ▶ scikit-learn: PCA; Importance of Feature Scaling.
- ▶ SciPy: statistical functions.
- ▶ Wattenberg, Viégas, Johnson (2016). How to Use t-SNE Effectively.

Ссылки, происхождение иллюстраций и журнал изменений — в источниках.