

Лекция-семинар

EDA, PCA и t-SNE

Покемоны и цифры

Лектор

Елизавета Власова

Практики

Кирилл Захаров

Камила Насибуллина

Михаил Подсытник

Игорь Толстокулаков

Распределения и качество данных

Центр, разброс, квартили и боксплот

Покемон как строка таблицы

Вид	ID	Основной тип	Второй тип	Вес, кг
Pikachu	25	electric	отсутствует	6
Snorlax	143	normal	отсутствует	460
Gengar	94	ghost	poison	40.5

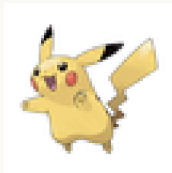
ID — идентификатор, хотя записан числом. Тип — категория.

Вес и боевые характеристики — числовые признаки.

Что изменится, если вычислить среднее по ID? Почему это не полезное описание покемона?

Типичный вес пяти покемонов

Pikachu
6 кг · 0.4 м



Snorlax
460 кг · 2.1 м



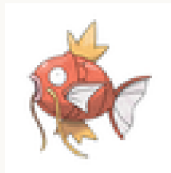
Onix
210 кг · 8.8 м



Gengar
40.5 кг · 1.5 м



Magikarp
10 кг · 0.9 м



Веса: 6; 460; 210; 40.5; 10 кг.

Вопрос: какой вес лучше описывает эту пятёрку? Предскажите, будут ли среднее и медиана близки.

Среднее и медиана той же пятёрки

Упорядоченные веса: 6; 10; **40.5**; 210; 460 кг.

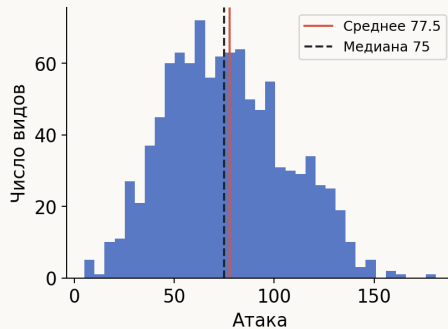
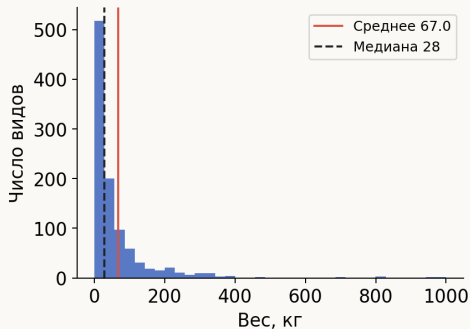
Медиана: 40.5 кг — центральное значение.

Среднее: $(6 + 10 + 40.5 + 210 + 460)/5 = 145.3$ кг.

Большой вес Snorlax и Onix сдвигает среднее. Медиана отвечает на вопрос о середине списка; среднее учитывает суммарный вес.

Теперь проверим, сохраняется ли различие на всей базе из 1025 видов.

Распределения веса и атаки



Вес: среднее 67.0 кг, медиана 28 кг. Одного среднего недостаточно.

Пропуски в Pokédex

Поле		Результат проверки
Восемь признаков	числовых	Пропусков нет
Второй тип		499 пустых значений: у вида один тип
Доля самцов		155 пустых значений: в схеме базы нет пола
capture_rate		Minior: «30 (Meteorite)255 (Core)»

Последнюю запись оставляем исходной и исключаем из числового PCA.

Как изменится смысл данных, если заполнить второй тип самым частым?

Квартили веса

Сортируем все 1025 вида по весу.

Граница	Где находится	В этой базе
Q_1	25-й процентиль	8.5 кг
Q_2	Медиана, 50-й процентиль	28 кг
Q_3	75-й процентиль	70 кг

Между Q_1 и Q_3 находится центральная половина распределения.

Вопрос: как измерить её ширину? Это и будет межквартильный размах.

Межквартильный размах и границы

$$IQR = Q_3 - Q_1 = 70 - 8.5 = 61.5 \text{ кг.}$$

Правило 1.5 IQR расширяет центральный интервал с обеих сторон:

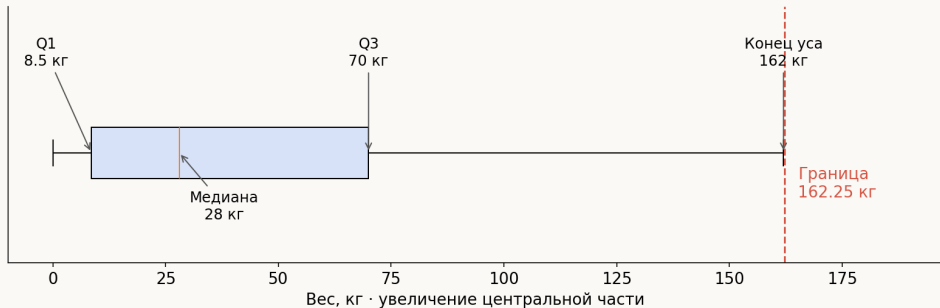
$$L = Q_1 - 1.5 IQR = -83.75 \text{ кг}, \quad U = Q_3 + 1.5 IQR = 162.25 \text{ кг.}$$

Вес ниже L или выше U отмечаем для проверки.

Отрицательная нижняя граница — результат правила, а не возможный вес.

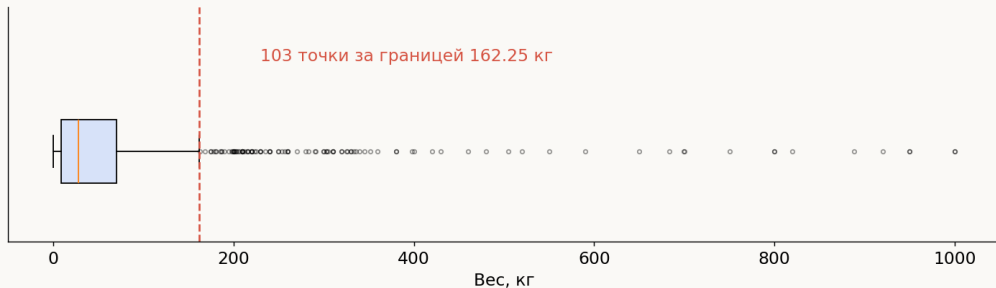
Будут ли отмечены Gengar (40.5 кг), Onix (210 кг) и Snorlax (460 кг)?

Как устроен боксплот



Коробка: Q1–Q3. Линия внутри: медиана. Усы доходят до крайних наблюдений внутри границ. Верхний ус здесь 162 кг, граница — 162.25 кг.

Боксплот всех весов



Отдельные точки за усами: 103 вида. Onix и Snorlax отмечены, Gengar — нет.
Боксплот выделяет необычные значения, но не доказывает ошибку данных.

Крайние значения веса

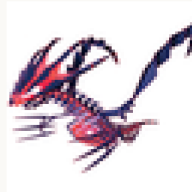
Cosmoem
999.9 кг · 0.1 м



Celesteela
999.9 кг · 9.2 м



Eternatus
950 кг · 20 м



Правило 1.5 IQR отмечает 103 вида по весу. Флаг требует проверки смысла.

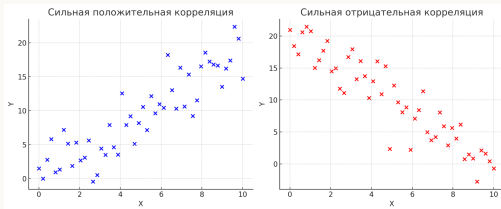
2

Связи признаков

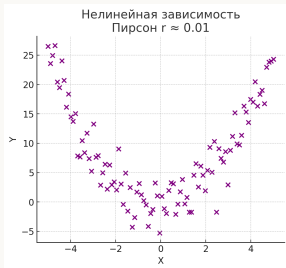
От вопроса к проверке на покемонах

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



Зависимость есть.

Прямая её не описывает.

Перед выводом по коэффициенту
смотрим на форму связи.

Вопросы о связях признаков

Гипотеза 1: более тяжёлые покемоны имеют более высокую атаку.

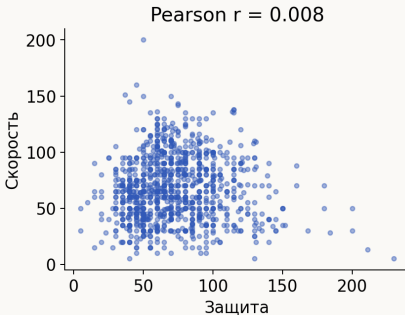
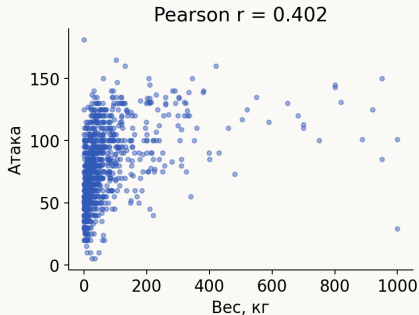
Гипотеза 2: более высокая защита связана с меньшей скоростью.

Как выглядели бы диаграммы рассеяния, если это так?

Какой знак корреляции мы ожидаем в каждом случае?

Сначала рисуем ожидаемое направление связи. Затем смотрим реальные точки.

Вес, атака, защита и скорость



Гипотеза «защита выше — скорость ниже» не подтверждается линейной корреляцией всей базы.

Сравнение двух групп

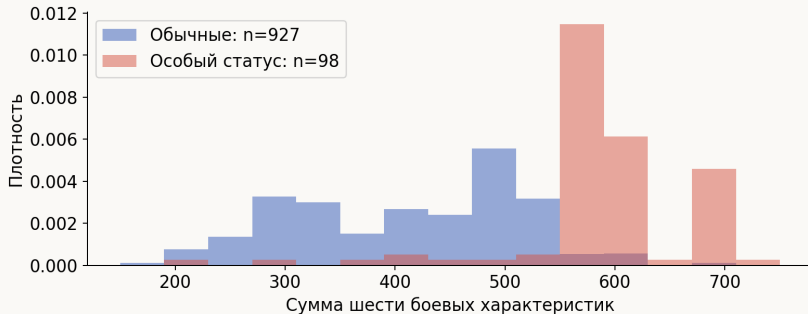
Вопрос: у покемонов с особым статусом выше сумма шести боевых характеристик?

Особый статус = хотя бы один флаг *legendary*, *sublegendary* или *mythical*.
Остальные виды образуют вторую группу.

Гипотеза: распределение первой группы сдвинуто к большим значениям.
Что проверяем: распределения, средние и медианы; также размер каждой группы.

Сначала опишите ожидаемую картинку. Сравниваем существующую базу видов.

Сравнение групп покемонов



Средняя сумма характеристик: 411.0 и 586.4. Особый статус объединяет три флага базы.

Что можно заключить о покемонах

Наблюдение

В этой базе у видов с особым статусом средняя сумма показателей выше на 175.4 пункта.

Описание

Показываем распределения и размеры групп, сравниваем медианы и средние.

Условия обобщения

Это созданные авторами игры виды, связанные эволюционными цепочками.

Обычный тест независимых случайных наблюдений нельзя применять автоматически.

Различие категорий не устанавливает причинный эффект статуса.

3

Стандартизация

Единицы измерения, дисперсия и соседи

Разброс веса внутри одной цепочки

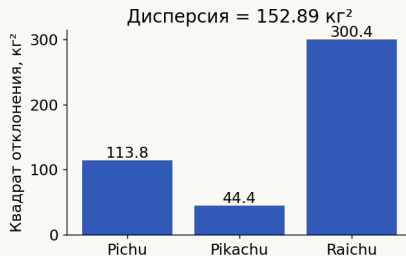
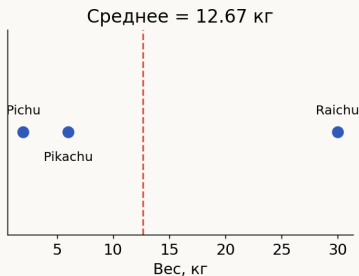
Покемон	Pichu	Pikachu	Raichu
Вес, кг	2	6	30

Среднее: $(2 + 6 + 30)/3 = 12.67$ кг.

Отклонения: -10.67 ; -6.67 ; $+17.33$ кг.

Отклонения со знаком в сумме дают ноль. Как измерить разброс, чтобы они не компенсировали друг друга?

Дисперсия на Pichu, Pikachu и Raichu



Квадраты отклонений усредняем: дисперсия 152.89 кг². СКО σ — её корень, 12.36 кг. Здесь делим на n , описывая эти три значения.

Меняем единицы — меняется дисперсия

Для той же тройки покемонов:

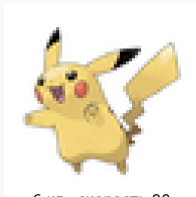
	Вес в кг	Вес в г
Среднее	12.67	12 666.67
СКО σ	12.36	12 364.82
Дисперсия	152.89	152 888 888.89

Отклонения и σ увеличились в 1000 раз. Их квадраты — в миллион.

Стандартизация $z = (x - \mu)/\sigma$ убирает эту зависимость: числитель и знаменатель меняют масштаб одинаково.

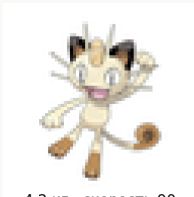
Пикачу и два кандидата

Pikachu



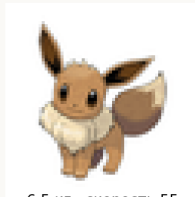
6 кг · скорость 90

Meowth



4.2 кг · скорость 90

Eevee



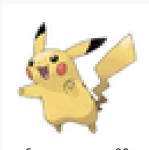
6.5 кг · скорость 55

Сравниваем **только вес и скорость**; выбираем **только между Meowth и Eevee**.
Кто ближе к Pikachu по евклидовому расстоянию?

Гипотеза: если перевести вес в граммы, выбранный сосед останется тем же.
Проверим?

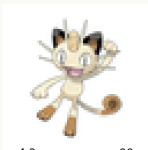
Расстояния, когда вес записан в кг

Pikachu



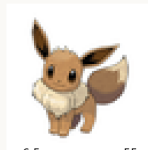
6 кг · скорость 90

Meowth



4.2 кг · скорость 90

Eevee



6.5 кг · скорость 55

Кандидат	Разница веса	Разница скорости	Расстояние
Meowth	1.8	0	$\sqrt{1.8^2 + 0^2} = 1.8$
Eevee	0.5	35	$\sqrt{0.5^2 + 35^2} \approx 35.0$

Meowth ближе: в исходных числах здесь решающей оказывается разница скорости.

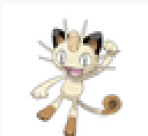
Те же покемоны, вес записан в г

Pikachu



6000 г · скорость 90

Meowth



4200 г · скорость 90

Eevee



6500 г · скорость 55

Кандидат	Разница веса	Разница скорости	Расстояние
Meowth	1800	0	1800
Eevee	500	35	$\sqrt{500^2 + 35^2} \approx 501.2$

Теперь ближе Eevee. Объекты прежние; мы изменили числовой вклад веса. Гипотеза не подтвердилась.

Стандартизация сохраняет выбор соседа

Для веса и скорости отдельно: вычитаем среднее и делим на σ .

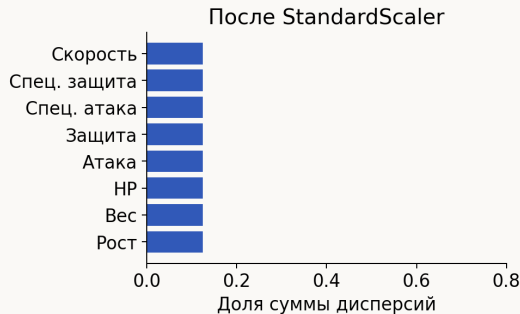
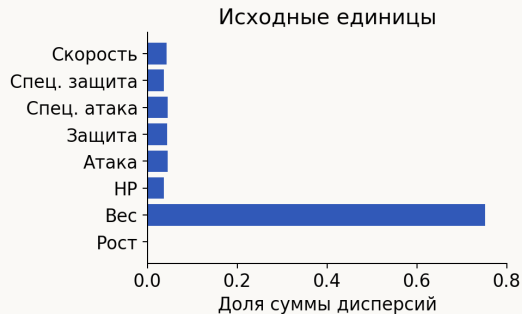
Параметры этого демонстрационного сравнения оценены по всей базе видов.

Кандидат	Вес до стандартизации: кг	Вес до стандартизации: г
Meowth	$d_z \approx 0.015$	$d_z \approx 0.015$
Eevee	$d_z \approx 1.22$	$d_z \approx 1.22$

И отклонение веса, и σ веса умножаются на 1000. Их отношение остаётся прежним.

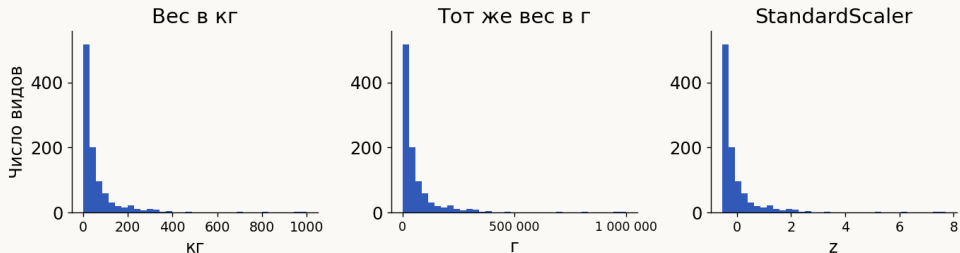
В обеих системах единиц выбирается Meowth. Мы сравниваем отклонения в единицах разброса.

Восемь признаков и их дисперсии



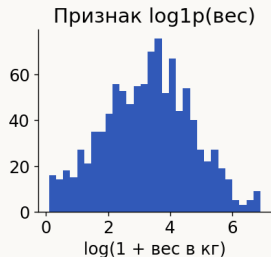
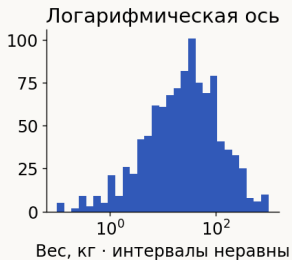
Каждая полоса = дисперсия признака / сумма восьми дисперсий. До: вес даёт 75.1%. После: у каждого дисперсия 1, доля 1/8. Вопрос: исчезнет ли правый хвост веса?

Линейная смена шкалы



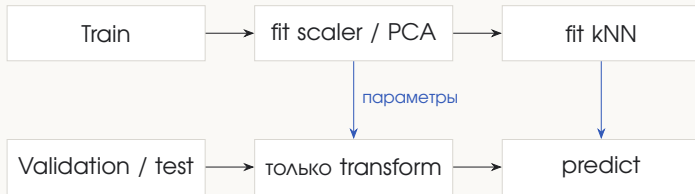
Проверка: при согласованных границах столбцы остаются той же формы. Меняются числа на оси. Гипотеза об исчезновении хвоста не подтвердилась.

Логарифмическая шкала и новый признак



В центре меняем отображение оси и интервалы гистограммы. Справа преобразуем сами значения для модели. Ни один вариант не гарантирует нормальность.

Обучение преобразований



Средние, масштабы и оси PCA определяем по train. На новых данных применяем то же отображение.

4

РСА: новые координаты

Проекции и потеря при уменьшении
размерности

Две защиты — одна координата

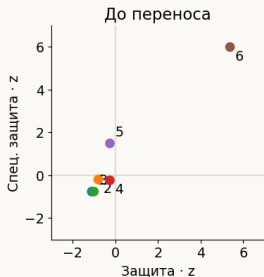
- 1 — Pikachu
- 2 — Bulbasaur
- 3 — Charmander
- 4 — Squirtle
- 5 — Snorlax
- 6 — Shuckle

Признаки: защита и спец.
защита; стандартизованы по
всей базе.



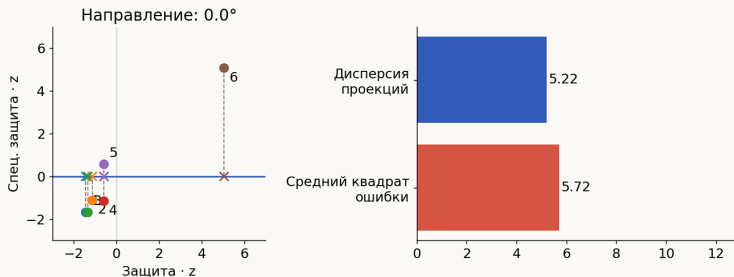
Как оставить одно число на покемона и потерять как можно меньше различий?

Переносим центр облака в ноль



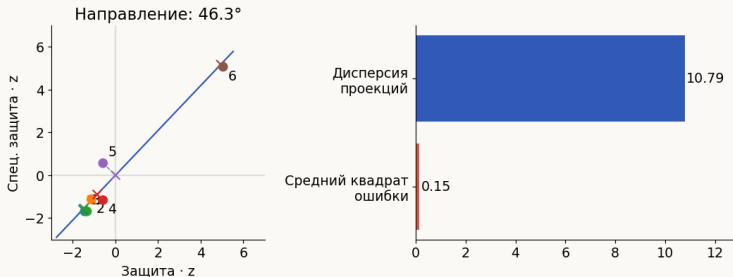
Вычитаем среднее шести точек. Цвет и номер обозначают того же покемона на всех стадиях. Расстояния сохраняются.

Первый вариант оси



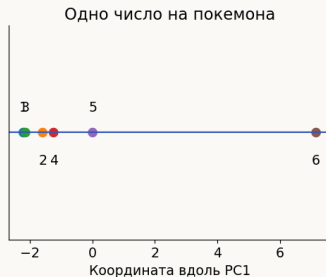
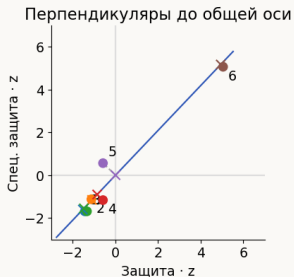
Оставим только горизонтальную координату. Красные отрезки показывают потерянную часть. Предскажите, поможет ли наклон оси вдоль облака.

Выбираем ось по разбросу проекций



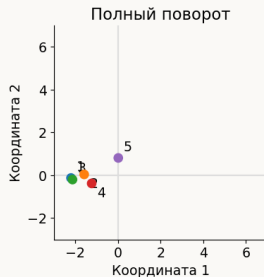
PCA выбирает направление 46.3° : дисперсия проекций максимальна, средний квадрат ошибки минимален.

Каждому покемону — одно число



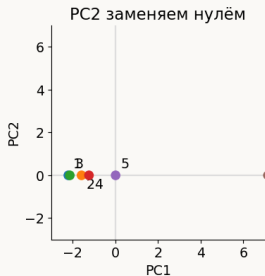
Проекция лежит на общей прямой. Её положение вдоль этой прямой — новая координата покемона.

Полный поворот обратим



Пока храним обе координаты, можно точно восстановить исходное положение каждого покемона.

Отбрасываем вторую координату



Одна ось сохраняет 98.7% дисперсии этой шестёрки. Красные отрезки — потерянные отклонения; большая доля не гарантирует полезность для любой задачи.

Две защиты покемона

Вопрос: можно ли заменить защиту и специальную защиту одним показателем?

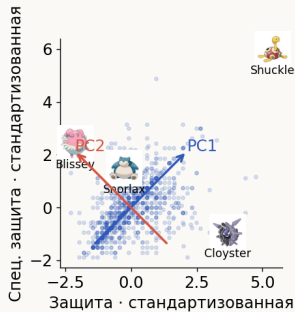
Гипотеза: они меняются согласованно, поэтому одна ось передаст большую часть различий.

Сначала стандартизируем оба признака.

На графике ищем: общее увеличение двух защит и отклонение в сторону одной из них.

Что потеряем у покемона, у которого одна защита высокая, а другая низкая?

РСА по двум признакам покемонов



PC1 описывает совместное изменение двух защит. PC2 — их дисбаланс. Удаление PC2 смешивает разные защитные профили.

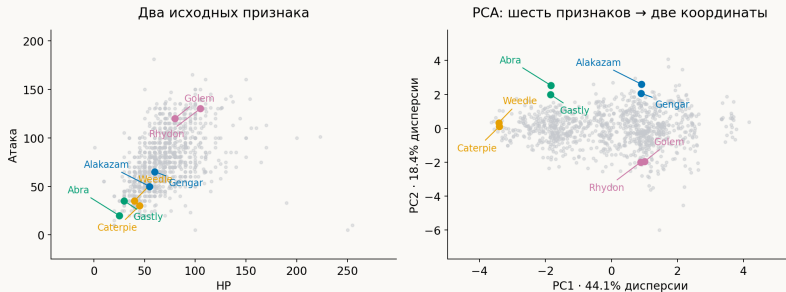
Шесть характеристик одного покемона

Вид	HP	Атака	Защита	Сп. ат.	Сп. защ.	Скорость
Abra	25	20	15	105	55	90
Caterpie	45	30	35	20	20	45
Gengar	60	65	60	130	75	110
Rhydon	105	130	120	45	45	40
Gastly	30	35	30	100	35	80
Alakazam	55	50	45	135	95	120
Weedle	40	35	30	20	20	50
Golem	80	120	130	55	65	45

Вся база: 1025 × 6. Стандартизация и PCA дают карту 1025 × 2.

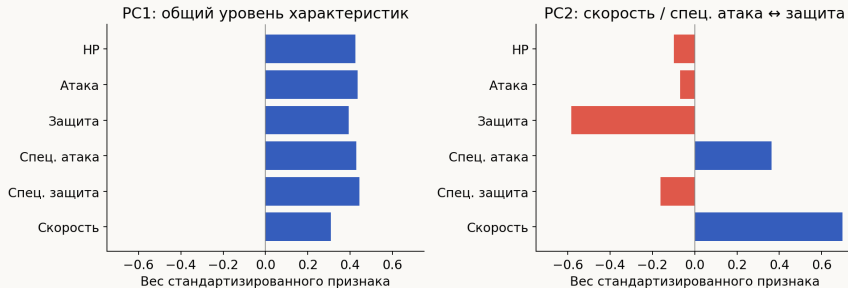
Кто похож на Abra: Caterpie, Gastly или Alakazam? Гипотеза: сходство профилей шести характеристик будет видно на карте.

Структура шести признаков на плоскости



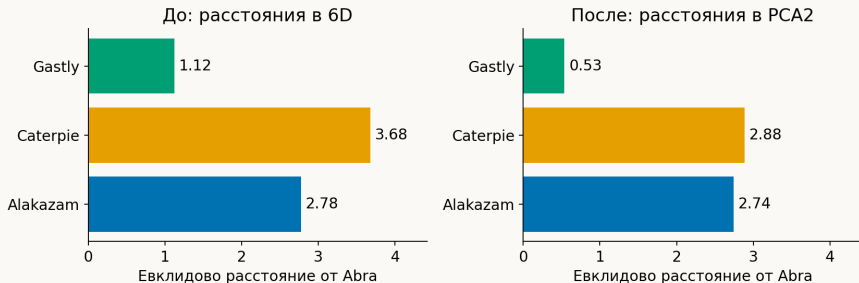
По HP и атаке Abra соседствует с Caterpie. На PCA видны различия профилей: Abra ближе к Gastly, а Alakazam и Gengar правее. Серым показана вся база; цветом выделены выбранные пары.

Смысл двух новых координат



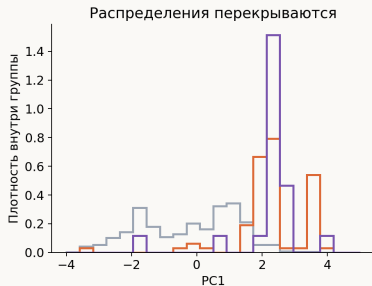
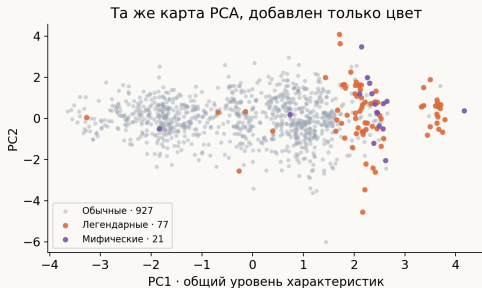
Правее — выше общий уровень характеристик. Выше — прежде всего больше скорость и спец. атака относительно защиты. Две координаты сохраняют 62.5% дисперсии шести признаков.

Проверка близости к Abra



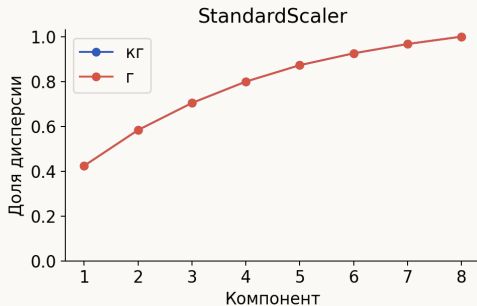
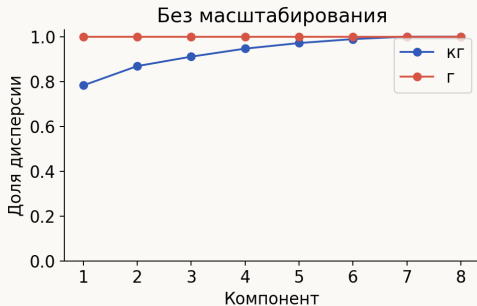
Из трёх кандидатов Gastly ближе всех и в 6D, и на карте. PCA делает сходство обозримым, но меняет расстояния: 37.5% дисперсии потеряно. Прогноз: где на этой карте окажутся легендарные?

Статус покемонов на карте PCA



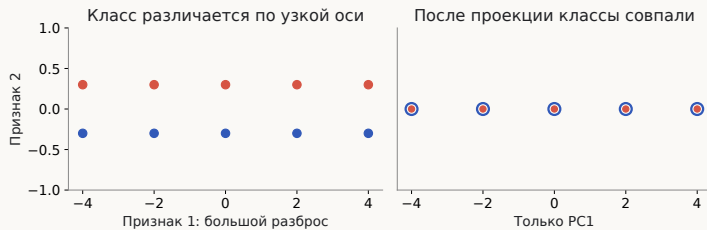
Легендарные чаще справа: общий уровень характеристик выше. Группы перекрываются. Статус не входил в PCA. Здесь legendary + sublegendary = 77; мифические показаны отдельно.

РСА зависит от единиц измерения



Восьмипризнаковая РСА без масштаба: РС1 сохраняет 78.2% в кг. При смене на г результат меняется.

Потеря полезного для задачи признака



- ▶ Полный ортогональный поворот сохраняет евклидовы расстояния.
- ▶ Отбрасывание компонент меняет расстояния и соседей.
- ▶ Большая сохранённая дисперсия ещё не гарантирует хороший прогноз.

5

Карты рукописных цифр

Одна картинка — одна точка в двух координатах

Одна картинка — одна точка

1797 изображений Digits; каждое — 8×8 яркостей.

Матрица входа: **1797 × 64**. Карта: **1797 × 2**.

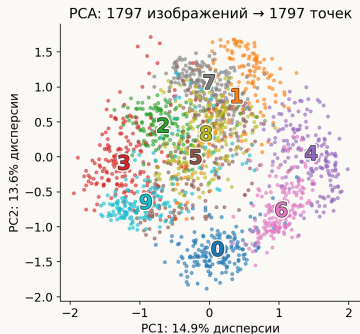
Яркости делим на 16: все признаки имеют одну шкалу (0, 1).

РСА обучаем на всех изображениях для описательной карты.

Метки 0–9 не передаём в РСА: используем их только для цвета.

Гипотеза: изображения одной цифры займут близкие области. Будут ли все десять групп разделены?

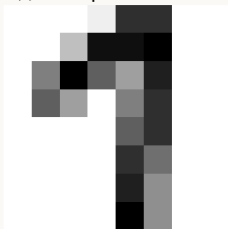
Все цифры в двух координатах PCA



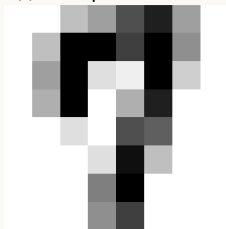
Есть области с преобладанием одного класса, но сильное перекрытие. Две компоненты сохраняют лишь 28.5% дисперсии. PCA не обучалось различать цифры.

Сходство конкретных 1 и 7

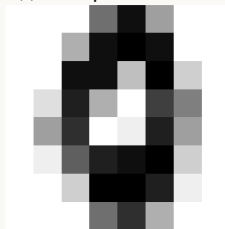
Метка 1 · индекс 1288
L2 до выбранной 1: 0.00



Метка 7 · индекс 480
L2 до выбранной 1: 1.79



Метка 0 · индекс 701
L2 до выбранной 1: 2.50



Показана ближайшая пара 1–7 по L2 в 64 пикселях; справа — ближайший к этой 1 ноль. $1.79 < 2.50$. Это пример сходства начертаний, не правило для всех цифр.

Сходство классов требует уточнения

Если сравнить **средние изображения классов** по L2:

Пара	Расстояние в 64 пикселях
1 и 7	1.95
1 и 0	2.63
1 и 8	1.31

По этому определению 1 ближе к 7, чем к 0. Но 7 — не самый близкий к 1 класс.

Всегда уточняем: сравниваем конкретные картинки, средние образы или соседства всех объектов?

6

t-SNE: карта соседств

Локальная структура и глобальные расстояния

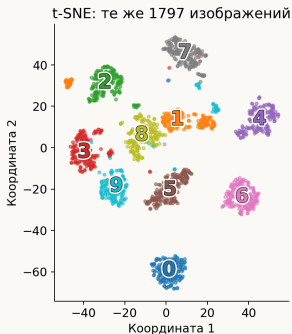
Что старается сохранить t-SNE

1. В 64 пикселях определяем, какие изображения являются близкими соседями.
2. Размещаем все 1797 точек на плоскости.
3. Меняем их положения, стараясь согласовать локальные соседства.

Это нелинейная карта: у неё нет двух общих направлений в пространстве пикселей, как у PCA.

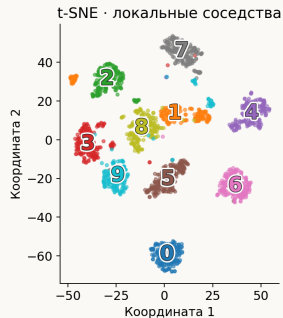
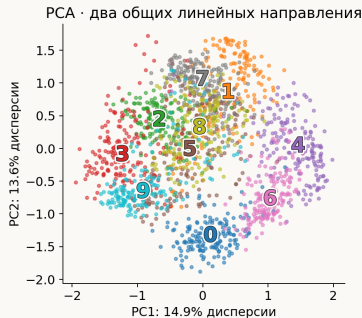
Гипотеза: классы разделятся заметнее. Но обязаны ли при этом сохраниться расстояния между группами?

Все цифры на карте t-SNE



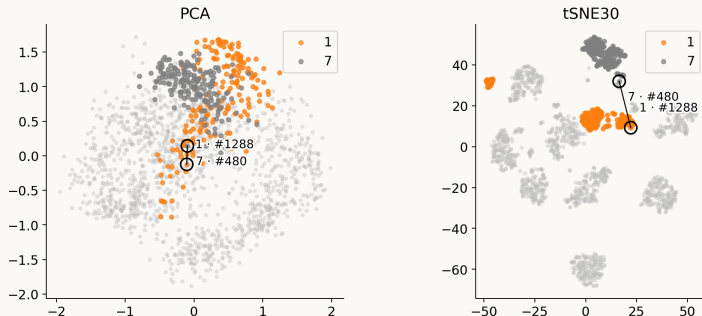
Вход тот же: 1797×64 . Выход: 1797×2 . $\text{perplexity}=30$, $\text{init}=\text{PCA}$, $\text{seed}=42$. Цвета раскрывают метки после обучения; алгоритм их не видел.

PCA и t-SNE на одних объектах



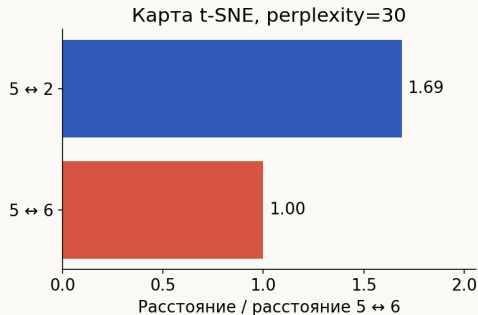
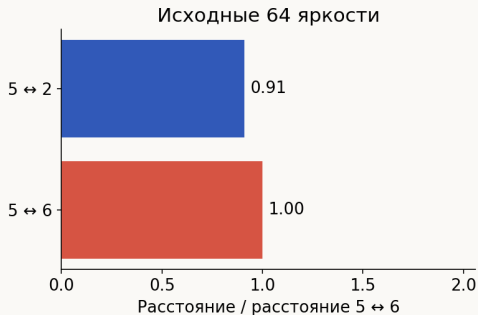
t-SNE сильнее разделяет многие классы. Это не означает, что промежутки между группами точнее передают сходство исходных изображений.

Похожие изображения на двух картах



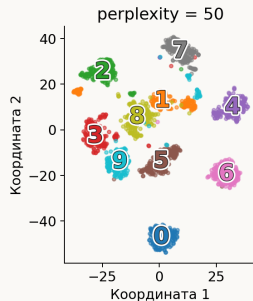
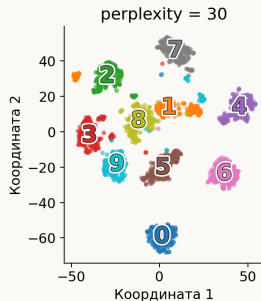
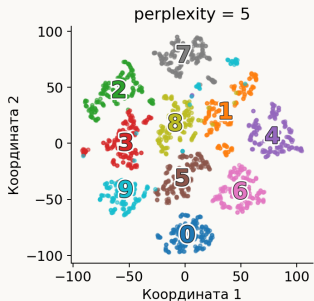
Обведены та же 1 и та же 7. Их исходные пиксели не менялись. Положение пары на двухмерной карте зависит от метода и соседств.

Проверяем относительные расстояния



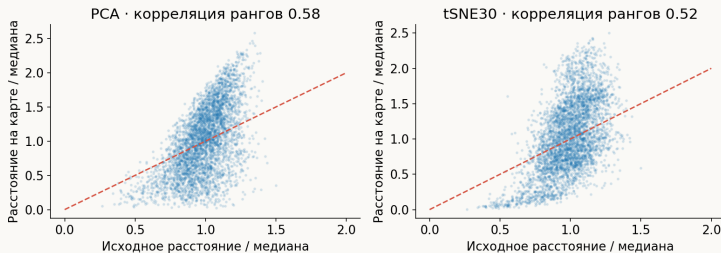
Сравниваем центры классов: в 64D расстояние 5–2 меньше 5–6 (отношение 0.91). На t-SNE оно больше (1.69). Общим масштабированием это не исправить.

Одна база, разные perplexity



perplexity задаёт масштаб рассматриваемого соседства. Меняются размеры и промежутки групп. По одному рисунку нельзя заключать, что один класс «вдвое ближе» другого.

Две координаты тоже искажают расстояния



Даже PCA с двумя компонентами теряет расстояния; точно сохраняет их только полный ортогональный поворот. t-SNE оптимизирует соседства, а не глобальные отношения расстояний.

Проверка в ноутбуке

1. Измените множитель IQR. Какие покемоны исчезают из списка отмеченных? Изменились ли сами данные?
2. Переведите кг в г. Покажите, почему меняется сосед и почему помогает стандартизация.
3. Выберите цифру и подсветите её на PCA и t-SNE. Где перекрываются классы?
4. Измените perplexity. Проверьте численно одно отношение расстояний до и после преобразования.

Отделяйте наблюдение на карте от утверждения об исходных объектах.

Теоретический минимум

- ▶ $IQR = Q_3 - Q_1$: ширина центральной половины распределения. Правило 1.5 IQR отмечает значения за границами.
- ▶ Боксплот: квартили, медиана, усы до наблюдений внутри границ, отдельные точки снаружи.
- ▶ Стандартизация учитывает масштаб дисперсии; форма распределения сохраняется.
- ▶ PCA даёт общие линейные направления. Отбрасывание координат создаёт потерю.
- ▶ На карте одна точка — один объект; цвет класса не обязательно участвовал в обучении.
- ▶ t-SNE согласует локальные соседства. Глобальные расстояния, размеры и промежутки групп читать буквально нельзя.

Ко(т)нец лекции

Связь со следующими темами

kNN и метрики расстояния.
Проверка качества на новых
данных. Представление объектов

определяет, кого алгоритм считает
похожим.

