Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 3: Обработка выбросов и нестандартных значений

Дата: 22.05.2025

Содержание лекции

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

- Что такое нестандартные значения? Почему в данных бывают выбросы?
- Виды выбросов
- Как определить выбросы? Многомерные и одномерные выбросы
- Одномерные методы
- Многомерные методы. Кластеризация
- Определили выброс, что делать дальше?

Виды аномалий: Выбросы и Новизна

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Два направления в анализе аномалий:

- Детектирование выбросов (Outlier Detection)

Обнаружение объектов, которые отличаются от обучающей выборки и уже в ней присутствуют

- Детектирование новизны (Novelty Detection)

Выявление новых объектов, которые ещё не встречались в обучающей выборке и появляются только в будущем

https://alexanderdyakonov.wordpress.com/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/

3

Причины выбросов

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Выбросы являются следствием:

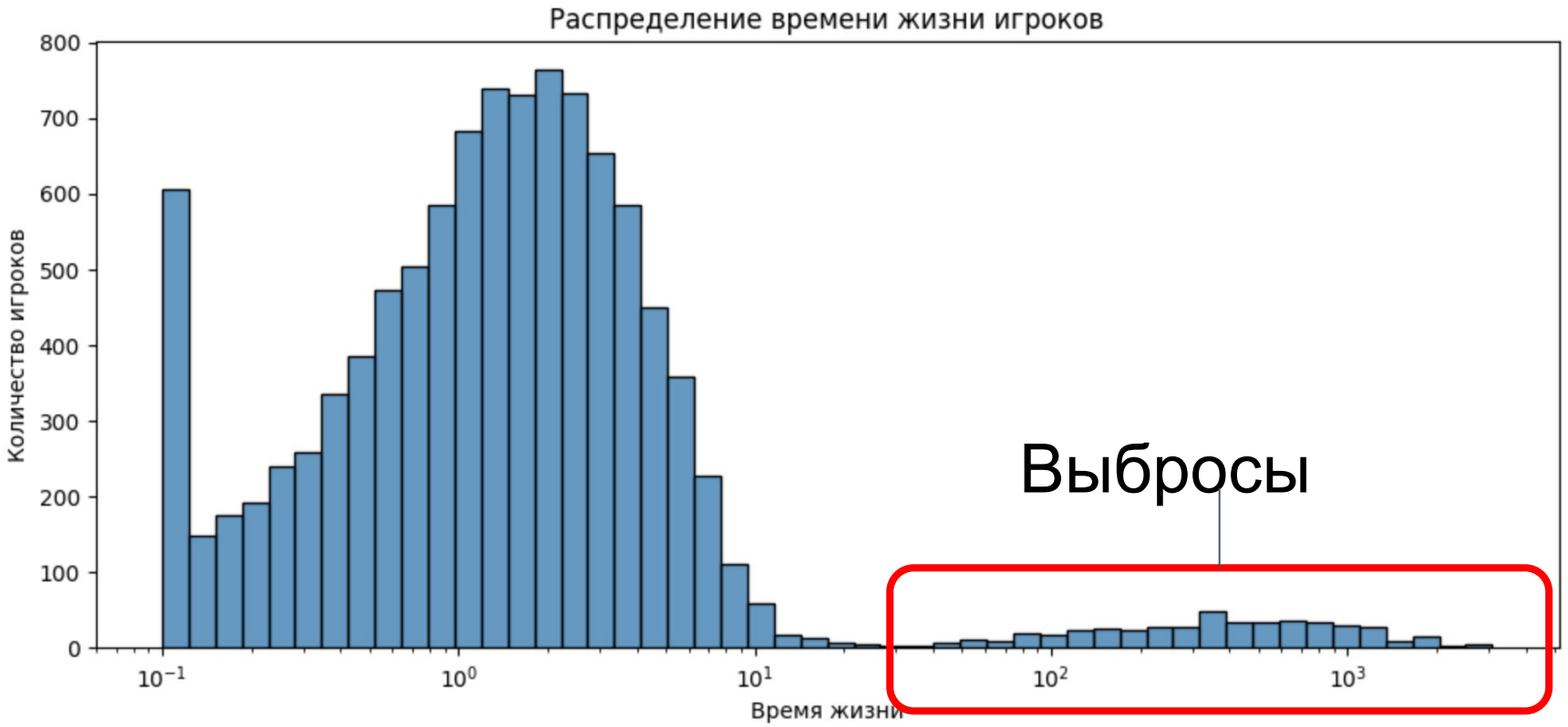
- ошибок в данных, ошибок агрегации или объединения нескольких источников данных, неверной записи
- неточности измерения
- присутствия объектов «других» распределений (например, показаниями сломавшегося датчика, другого типа клиентов)
- редких, но реальных событий или аномалий (например, экстремальных погодных условий)
- изменений в поведении системы или объекта наблюдения (например, смены модели работы оборудования)
- случайных шумов в процессе сбора данных
- ошибок при трансформации или обработке данных (например, неправильное масштабирование, сдвиг)
- влияния внешних факторов, не учтённых в модели или сборе данных (например, вмешательство человека)

Почему выброс - это проблема?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

время жизни игрока в игре: большая доля игроков удаляет игру сразу, как установили, но есть часть игроков, которые играют много.

Среднее без выбросов = 1.95, среднее с выбросами - 25.8



1. Выбросы искажают статистики:

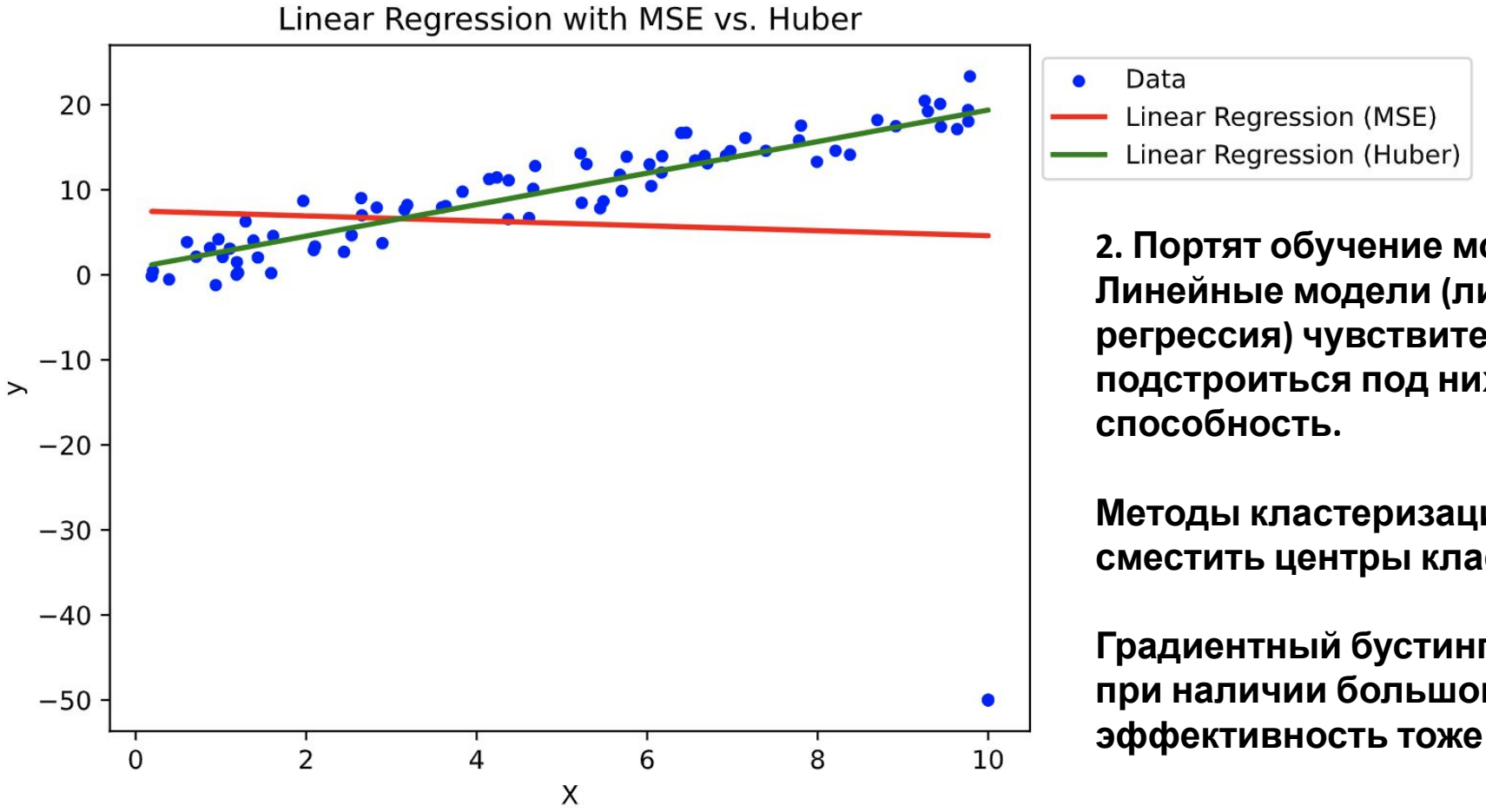
- **Среднее значение сильно смещается в сторону выбросов.**
- **Стандартное отклонение и дисперсия становятся завышенными.**

Корреляции могут быть искажены.

Почему выброс - это проблема?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Выбросы смещают и уравнение регрессии



2. Портят обучение моделей
Линейные модели (линейная регрессия, логистическая регрессия) чувствительны к выбросам: модель может подстроиться под них и потерять обобщающую способность.

Методы кластеризации (например, k-means) могут сместить центры кластеров.

Градиентный бустинг и деревья более устойчивы, но при наличии большого числа выбросов их эффективность тоже падает.

Как обнаружить выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Метод Z-оценки используется для обнаружения выбросов в данных, предполагая, что признак имеет приближённо нормальное распределение.

Z-оценка вычисляется по формуле:

$$z = \frac{x - \mu}{\sigma}$$

Где:

- x — значение признака,
- μ — среднее значение признака,
- σ — стандартное отклонение признака.

Пороговое правило для определения выбросов: $|z| > 3$

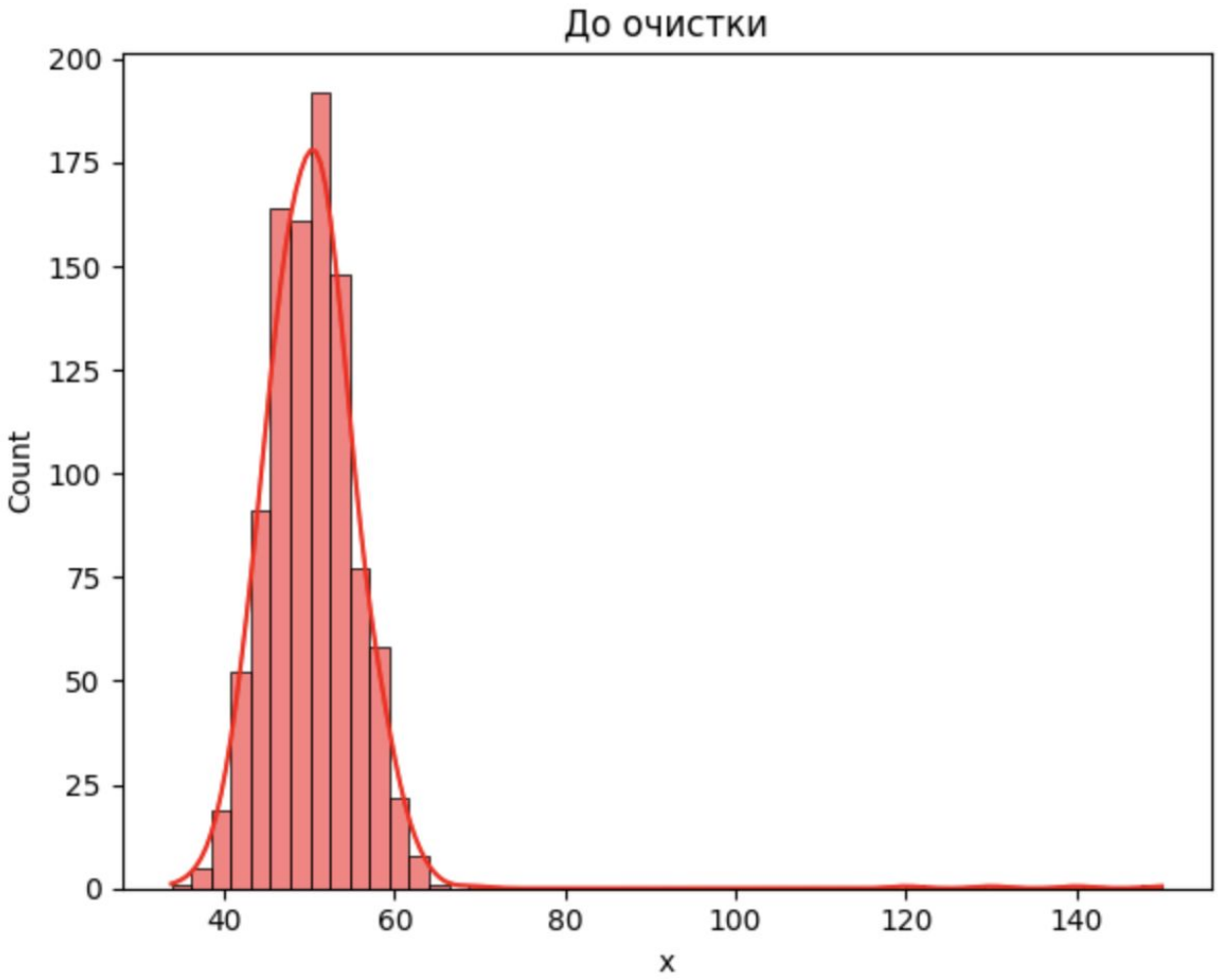
Это означает, что значения, отстоящие от среднего более чем на три стандартных отклонения, считаются выбросами.

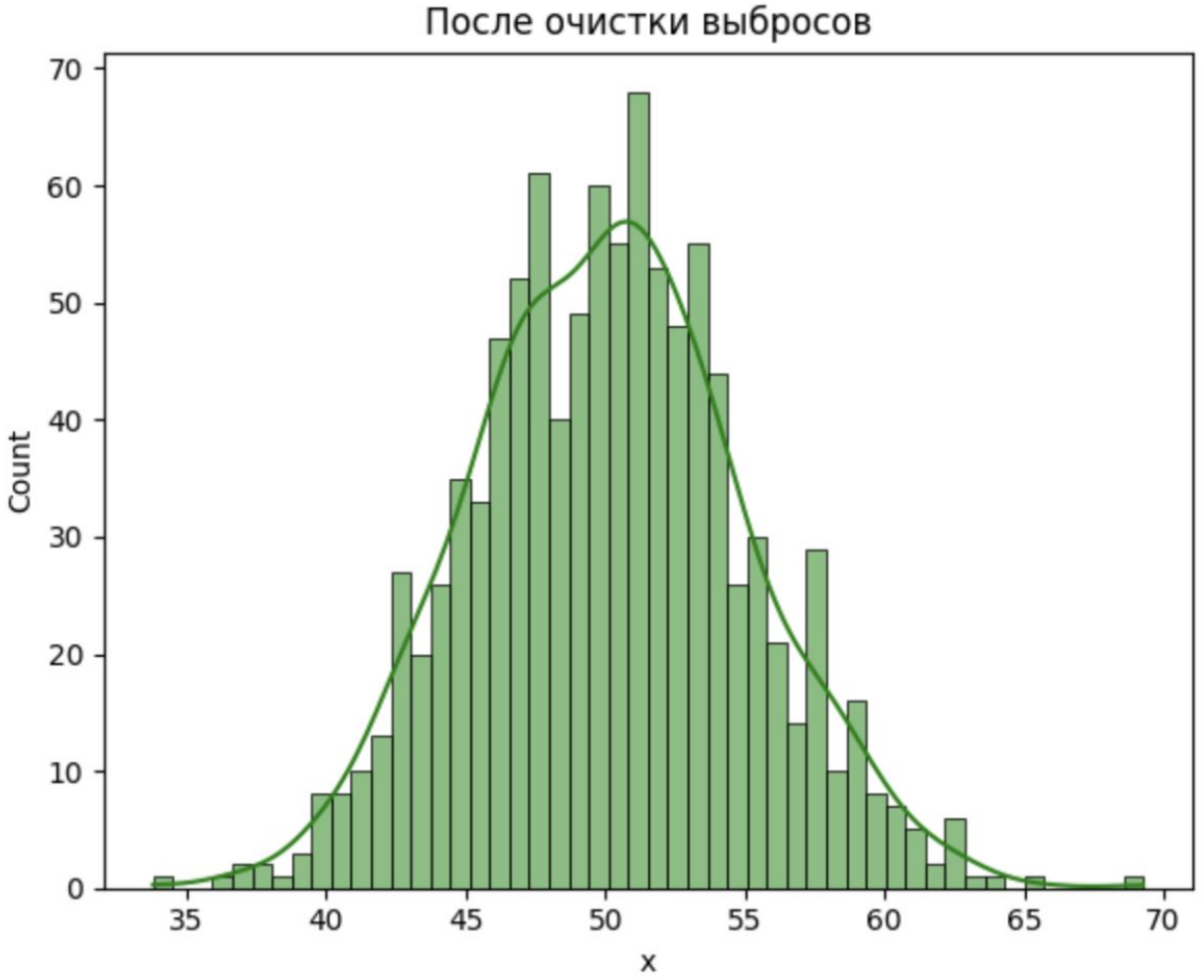
Важно: метод чувствителен к распределению. Он работает корректно только в случае, если данные приблизительно нормальны. Если распределение асимметрично или имеет тяжёлые хвосты, Z-оценка может давать ложные срабатывания.

Как обнаружить выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Метод Z-оценок





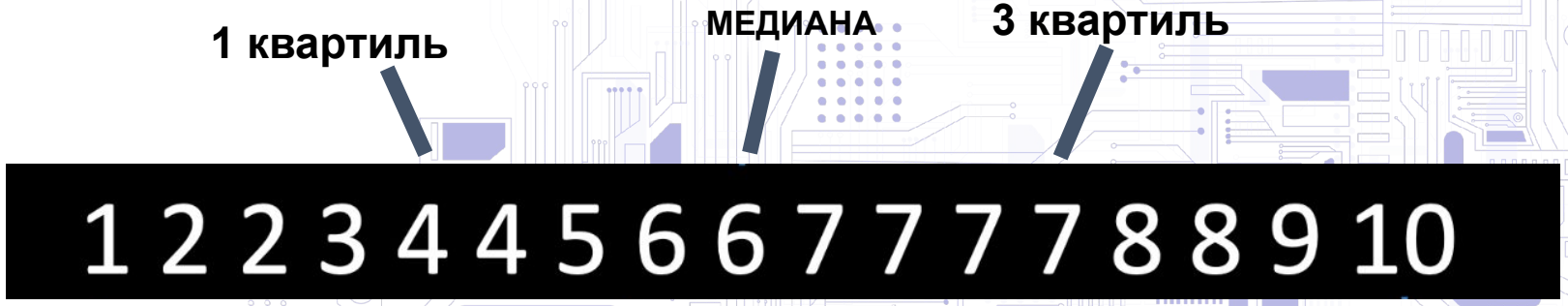
Как обнаружить выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Что такое квантиль? Квантиль задает значение, ниже которого находится определенная доля данных в распределении.

1) отсортируем ряд
2) разделим ряд на две части. Точка, которой мы делим - это и есть медиана. То есть ниже медианы 50% данных
3) Если мы ряд разделим на 3 равные части - то получим квартили. Ниже 1 квартиля 25% всех данных, ниже 2 квартиля - 50%, ниже 3 квартиля 75%

Получается, Квантиль задает значение, ниже которого находится определенная доля данных в распределении.
Медиана = 0.5 квантиль = 50% персентиль = 2 квартиль - это значение,



Как обнаружить выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

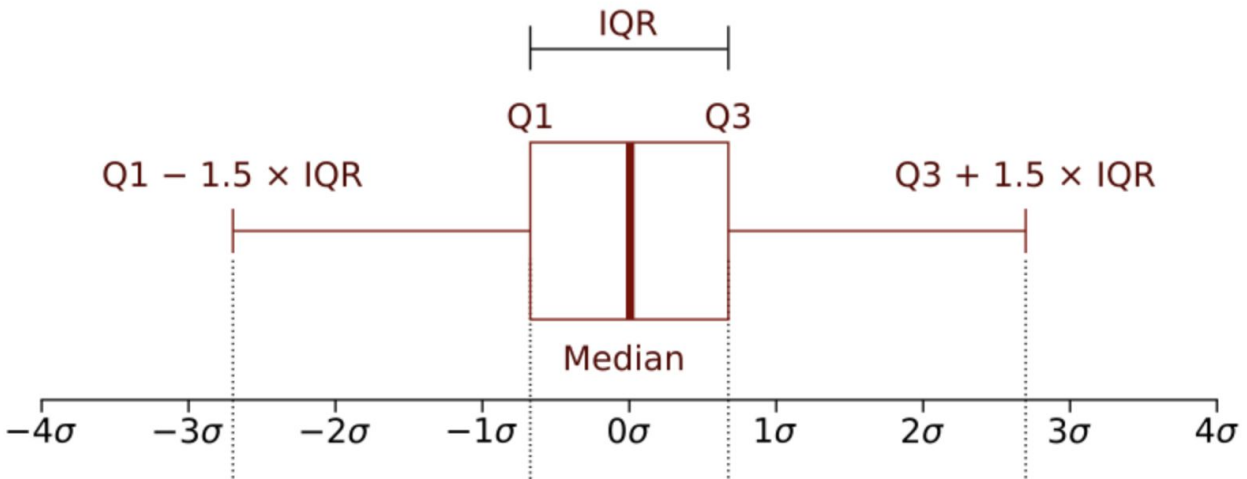
Одномерные методы обнаружения выбросов: Используются, когда анализируем один признак (фичу) отдельно.

Выбросы — это значения, выходящие за пределы диапазона:

$$[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$$

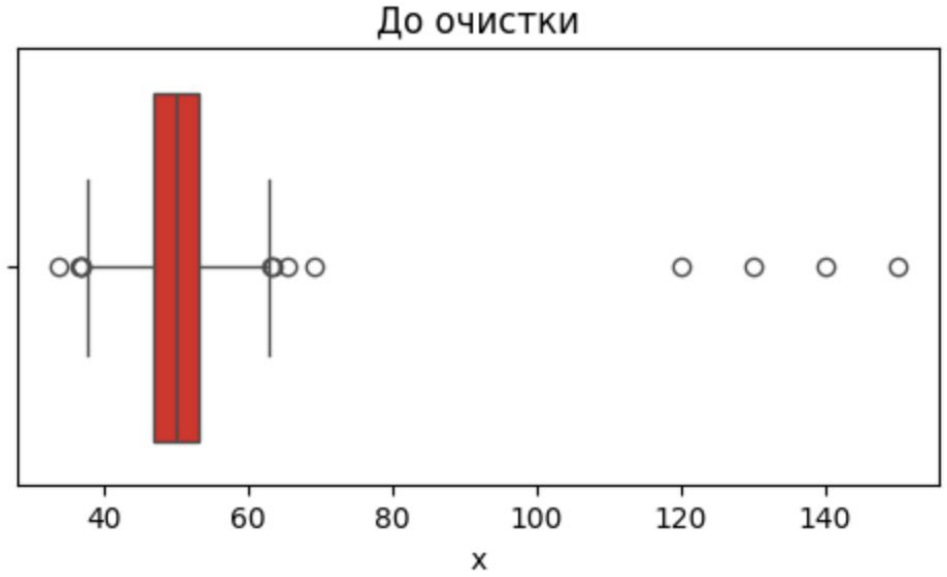
Где:

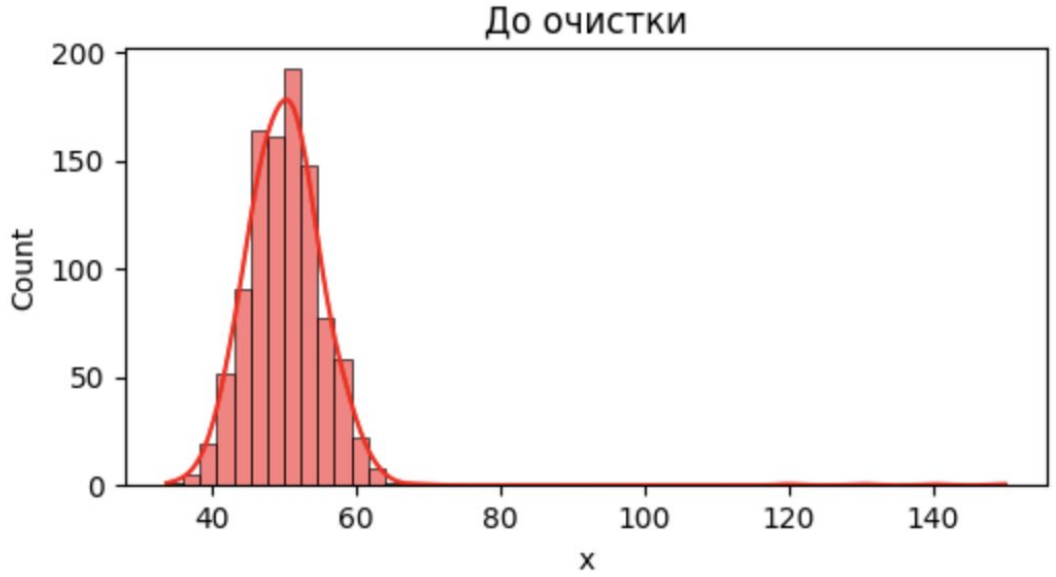
- Q_1 – 25-й перцентиль,
- Q_3 – 75-й перцентиль,
- $IQR = Q_3 - Q_1$ – межквартильный размах.

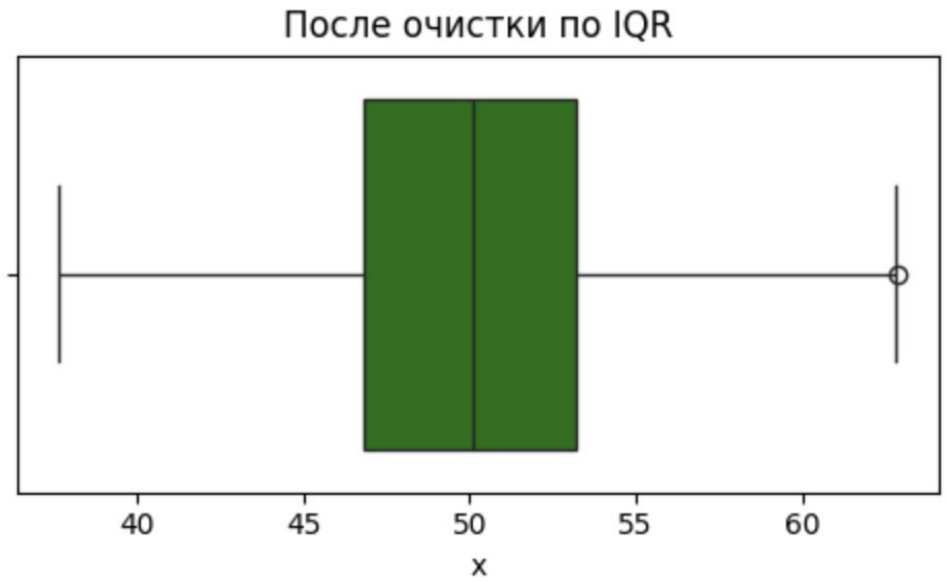


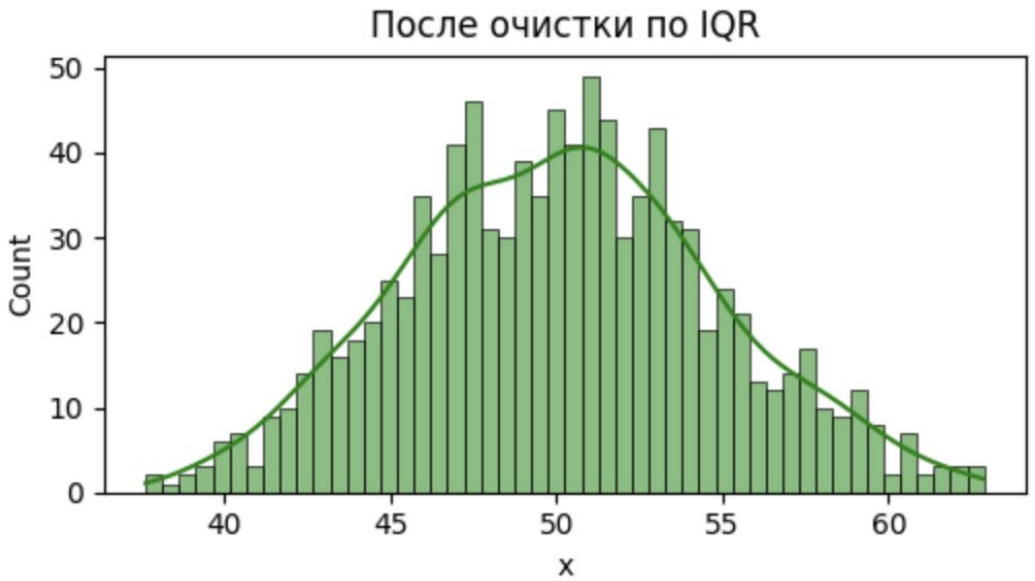
Как обнаружить выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС





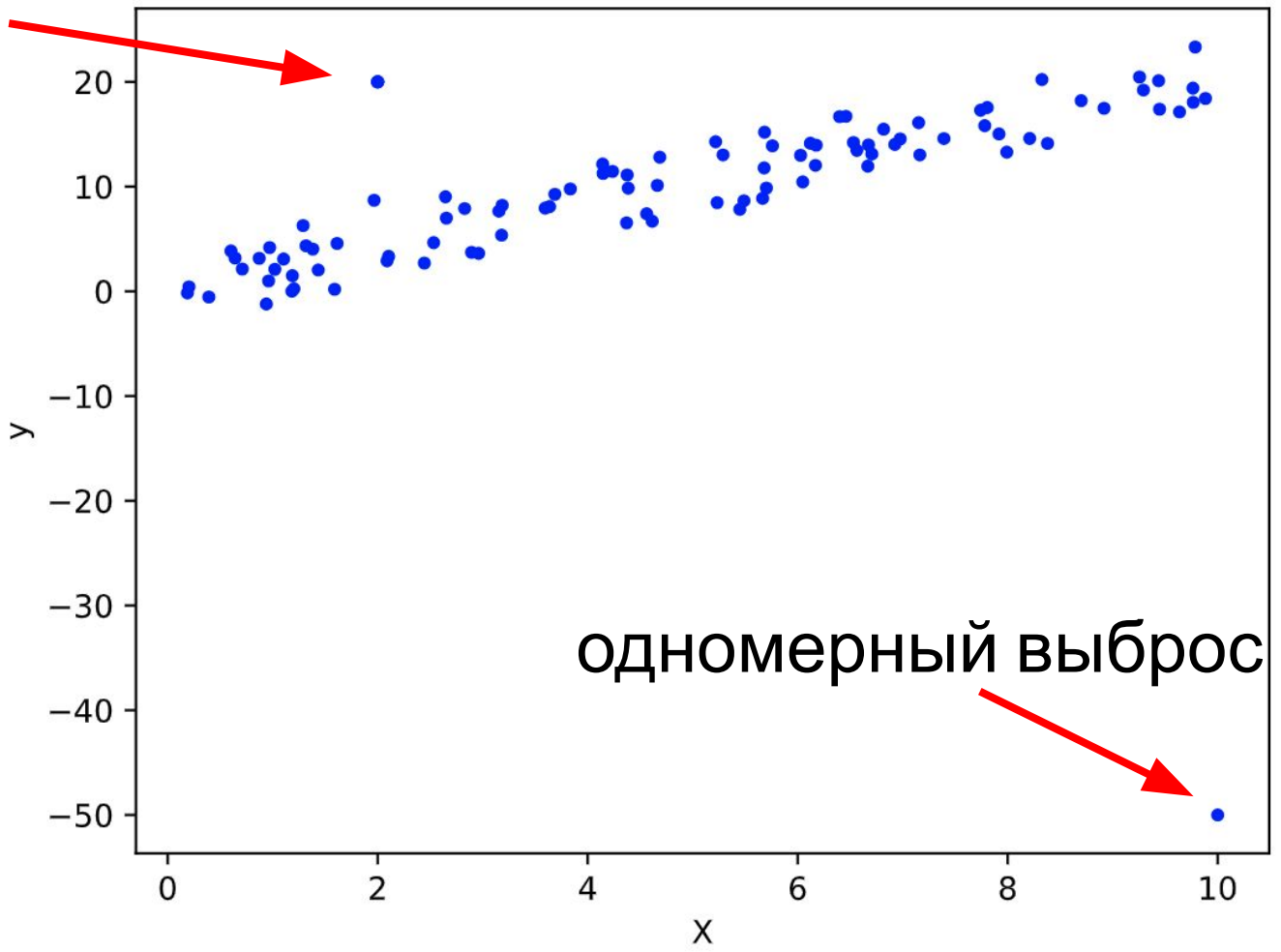




Многомерные и одномерные выбросы

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

многомерный выброс



Как обнаружить многомерные выбросы?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Многомерные выбросы невозможно обнаружить, используя одномерные методы.

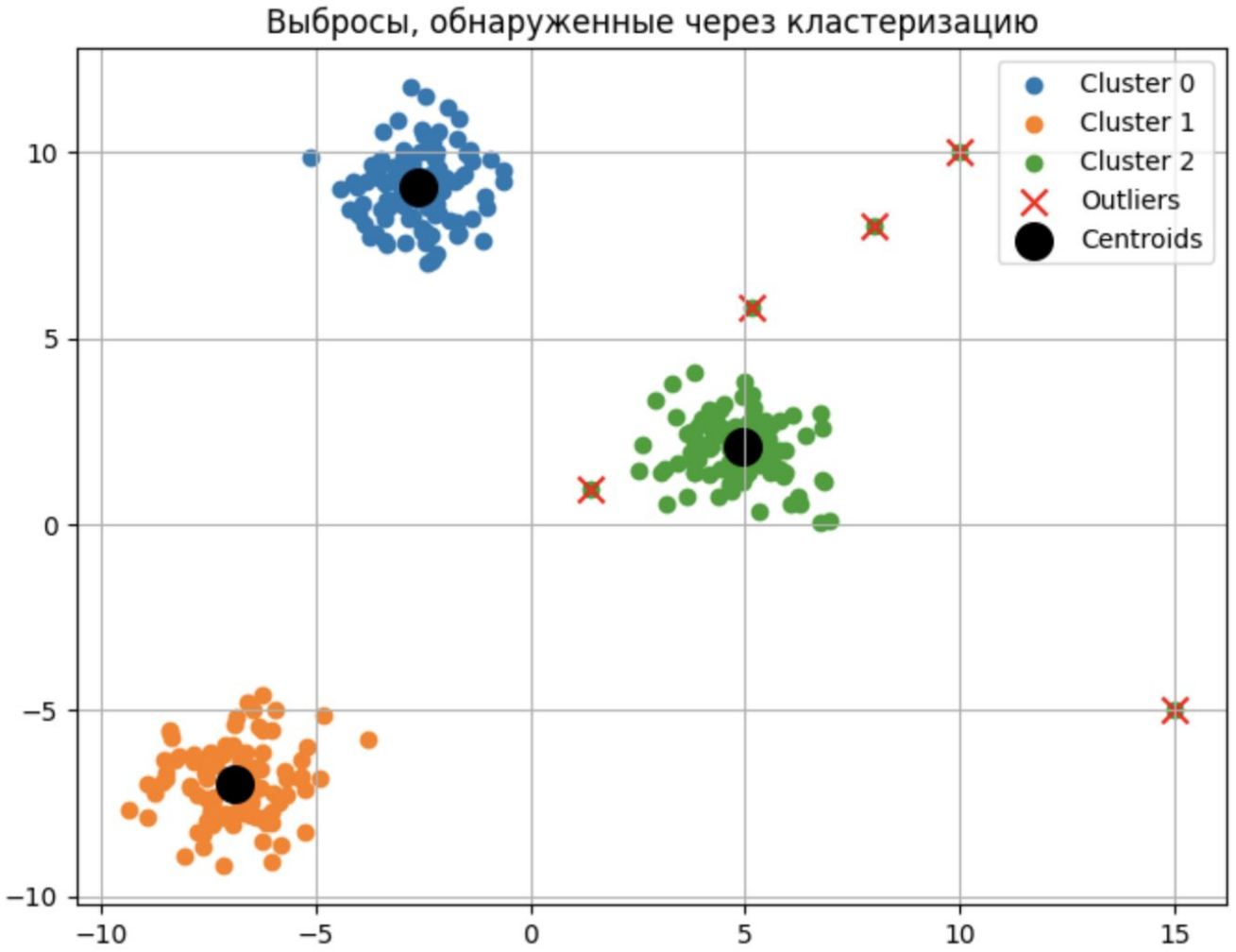
Используют следующие методы:

- методы кластеризации
- метод Isolation Forest (изучим на лекциях по деревьям)
  - Строит случайные деревья разделения.
  - Выбросы быстрее изолируются → имеют меньшую "глубину"
- One-Class SVM
  - Обучается на "нормальных" данных, затем выявляет отклонения
- Другие методы машинного обучения

Как кластеризация помогает определить выброс

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

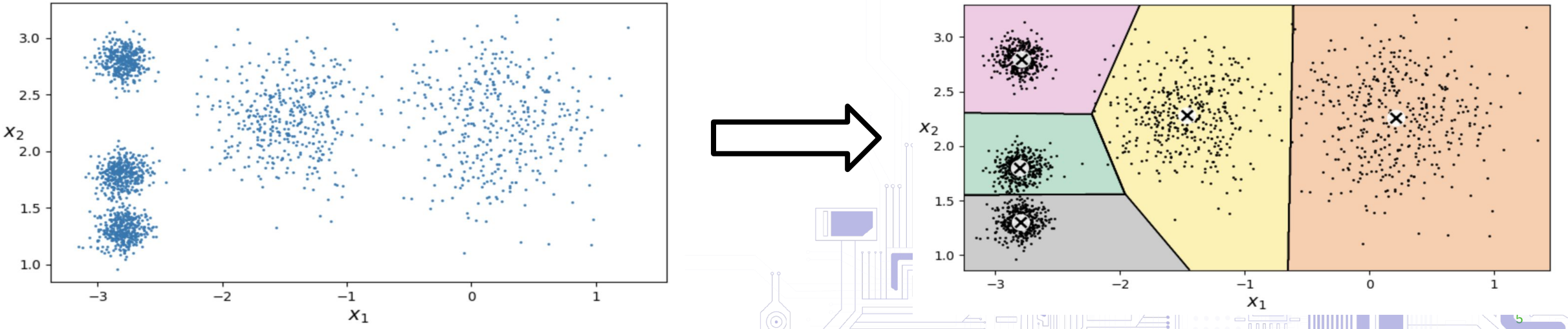
выявление многомерных выбросов с помощью кластеризации



Что такое кластеризация?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Кластеризация (англ. cluster analysis) — задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

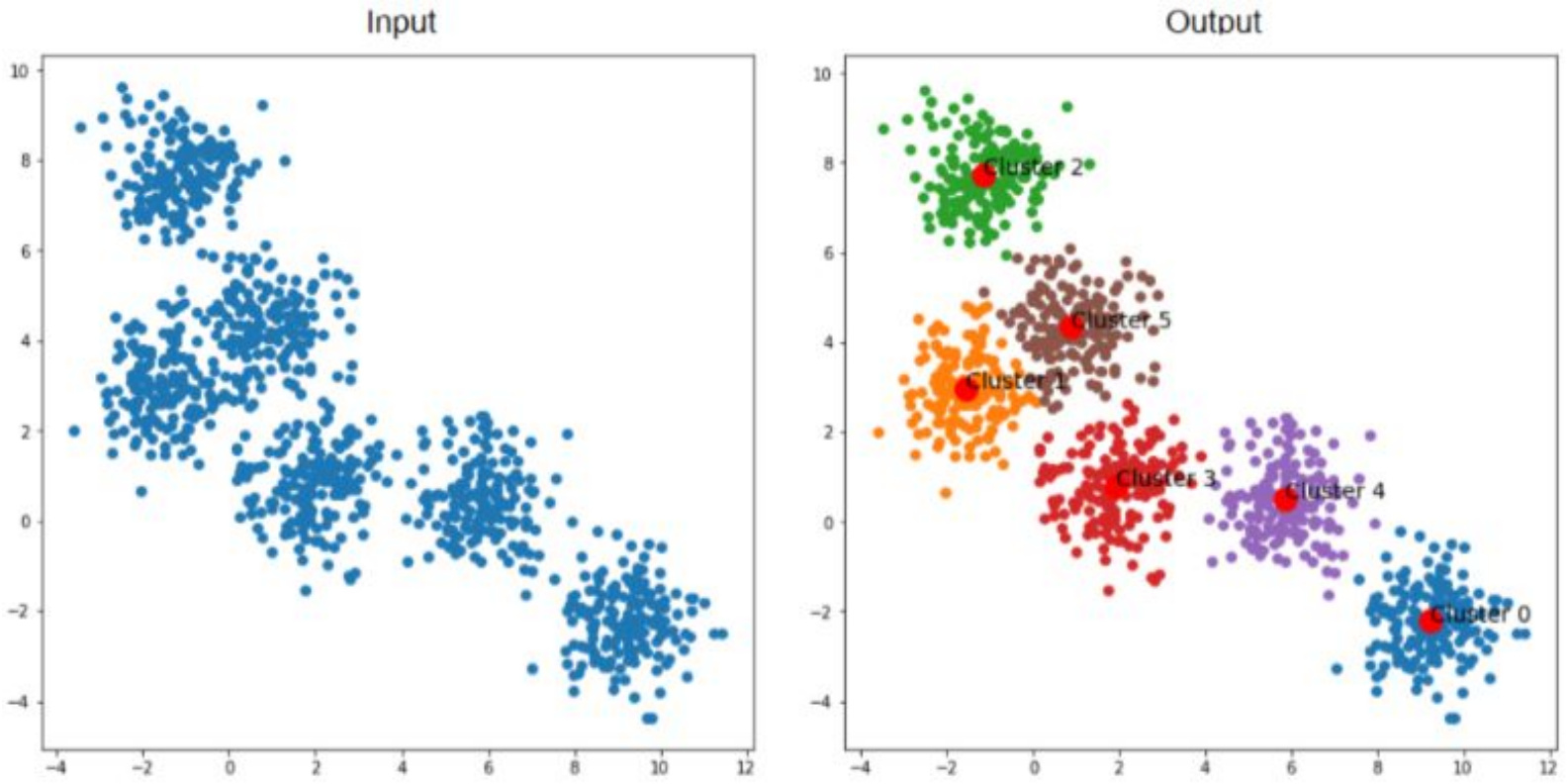


Что такое кластеризация?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

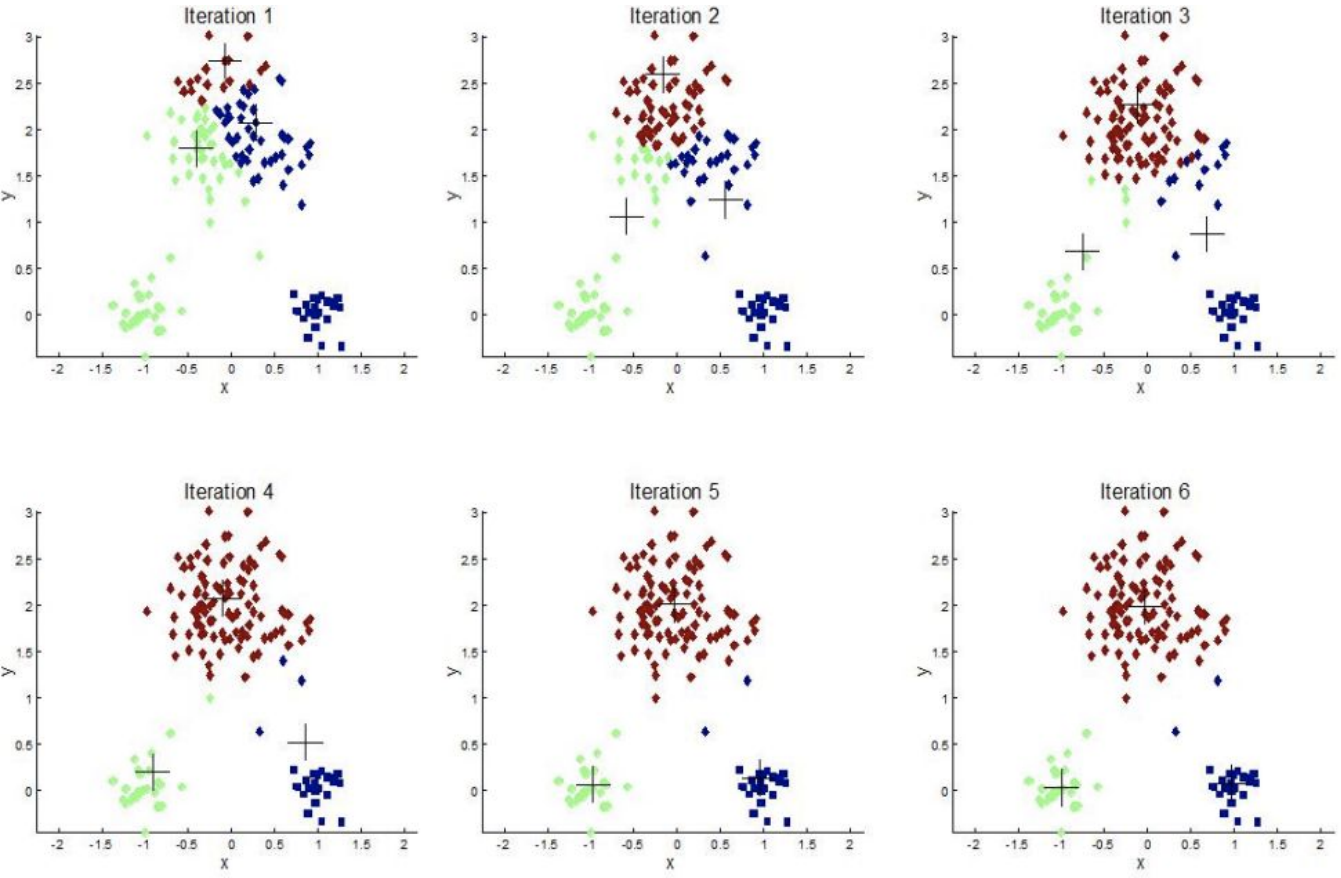
- разбиваем объекты на конечное множество классов
- нет понимания, какой будет природа этих классов
- то, что модель кластеризации какие-то объекты сочла «похожими», отнесся к одному классу, будет новой для нас информацией

Кластеризация — это задача обучения без учителя (unsupervised classification)



Кластеризация kmeans

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС



1. Случайным образом расставляются k-центров
2. Каждое наблюдение относится к тому кластеру, к центру которого оно ближе всего
3. Каждое наблюдение попадает только в один кластер
4. Пересчитывается центр каждого кластера
5. Шаги 2 и 3 повторяются,пока кластеры не перестанут изменяться

Кластеризация kmeans

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Алгоритм стремится минимизировать среднеквадратичное отклонение от центра для элементов каждого кластера.

$$R = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2$$

k - число кластеров

C_i - полученные кластеры

c_i - центр i-го кластера

Кластеризация kmeans

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Чтобы сравнить два объекта, необходимо иметь критерий, на основании которого будет происходить сравнение. Критерий - расстояние между объектами (метрика)

- $d(\vec{X}_i, \vec{X}_j) \geq 0$
- $d(\vec{X}_i, \vec{X}_j) = 0, \rightarrow \vec{X}_i = \vec{X}_j$
- $d(\vec{X}_i, \vec{X}_j) = d(\vec{X}_j, \vec{X}_i)$
- $d(\vec{X}_i, \vec{X}_k) \leq d(\vec{X}_i, \vec{X}_j) + d(\vec{X}_j, \vec{X}_k)$

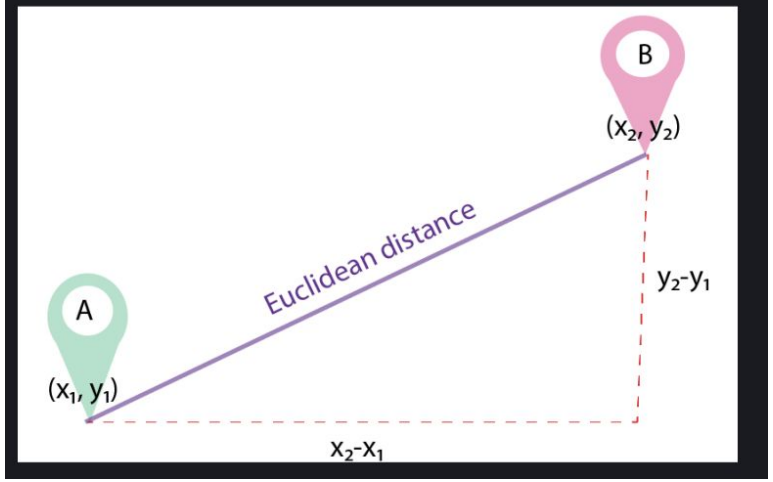
Гиперпараметры алгоритма kmeans:

- количество кластеров
- метрика расстояния

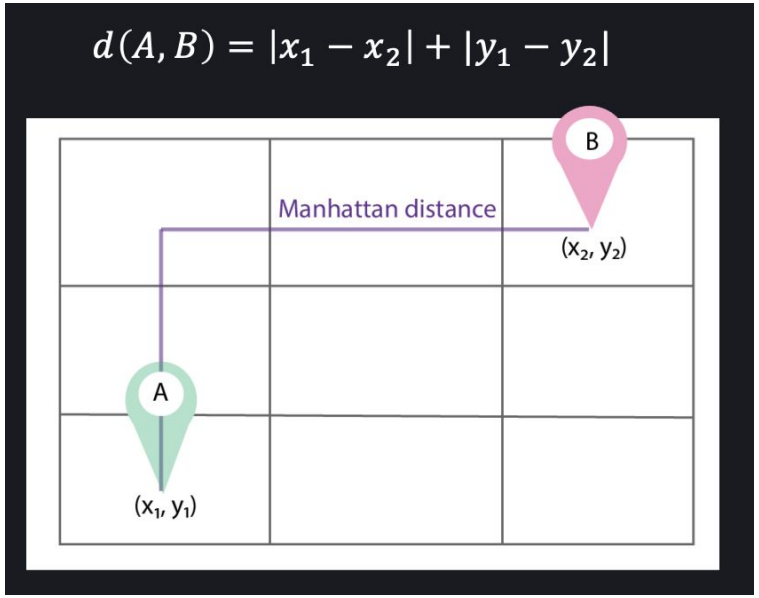
Кластеризация kmeans. Расстояние

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

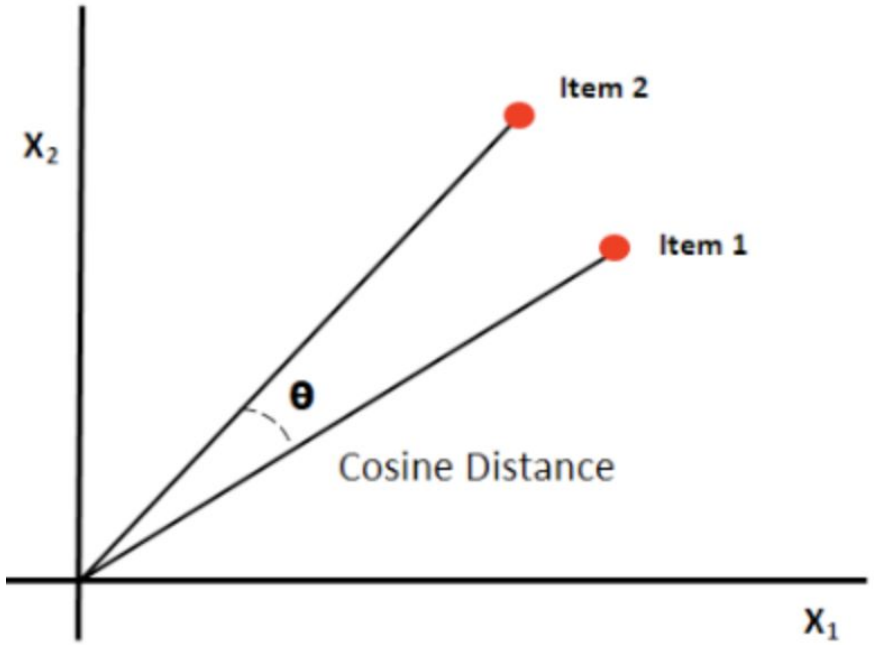


$$d(A,B) = |x_1 - x_2| + |y_1 - y_2|$$



косинусная мера — это функция близости, а не расстояние, так что чем больше её значения, тем ближе друг к другу векторы.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



Кластеризация kmeans. Плюсы алгоритма

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Простота и понятность: K-means является относительно простым и легко понимаемым алгоритмом. Это делает его привлекательным для использования и внедрения в различных областях

Высокая эффективность для сфер сферических кластеров: В случае, если кластеры имеют приблизительно сферическую форму и примерно одинаковый размер, k-средних может давать хорошие результаты

Линейная сложность в среднем случае: В среднем случае алгоритм имеет линейную сложность, что делает его относительно эффективным для средних размеров данных

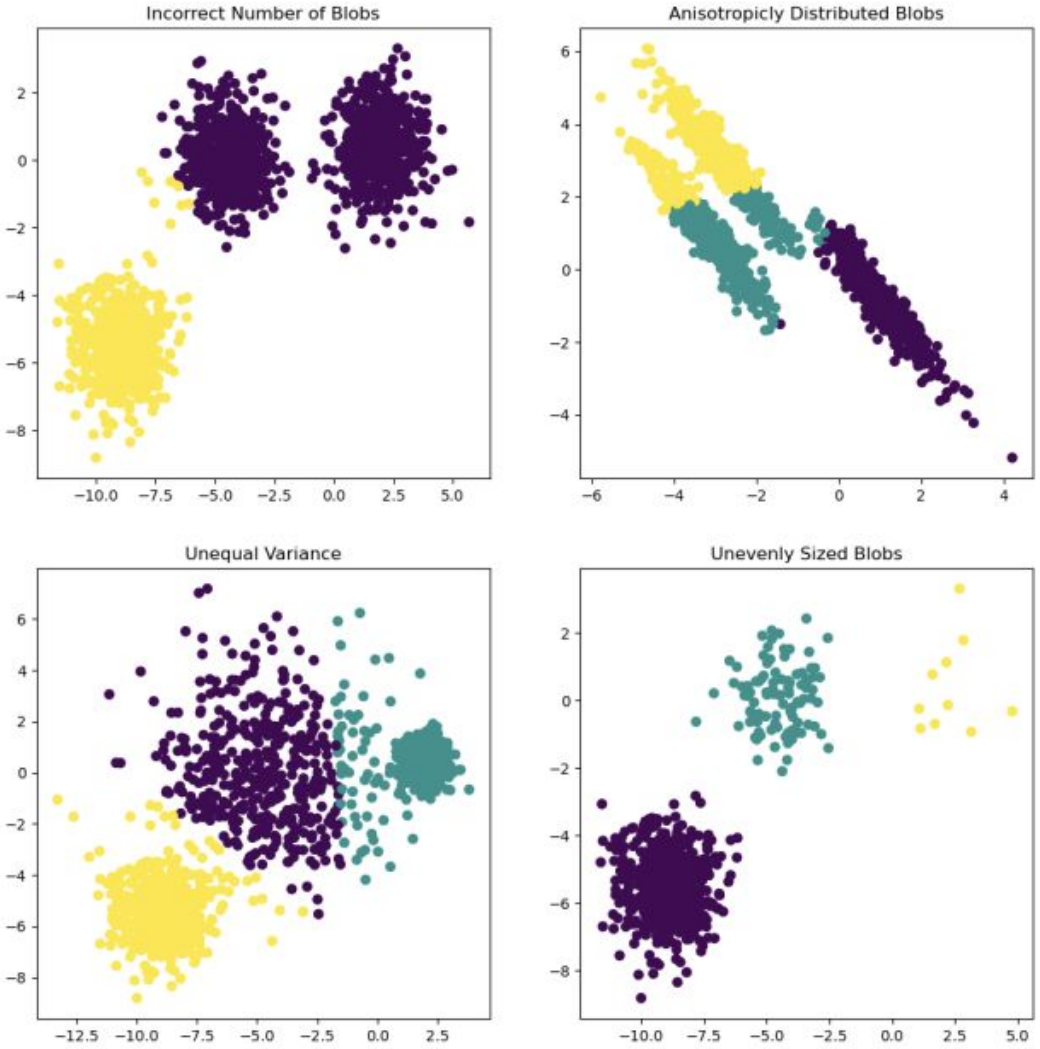
Интерпретируемость результатов: Результаты кластеризации методом k-средних обычно легко интерпретировать, особенно когда кластеры имеют четкие центроиды

Кластеризация kmeans. Минусы алгоритма

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Алгоритм может выдавать контринтуитивные результаты, если:

- Указано не то число кластеров
- Кластеры не выпуклые и близко расположены
- Разная дисперсия близких кластеров



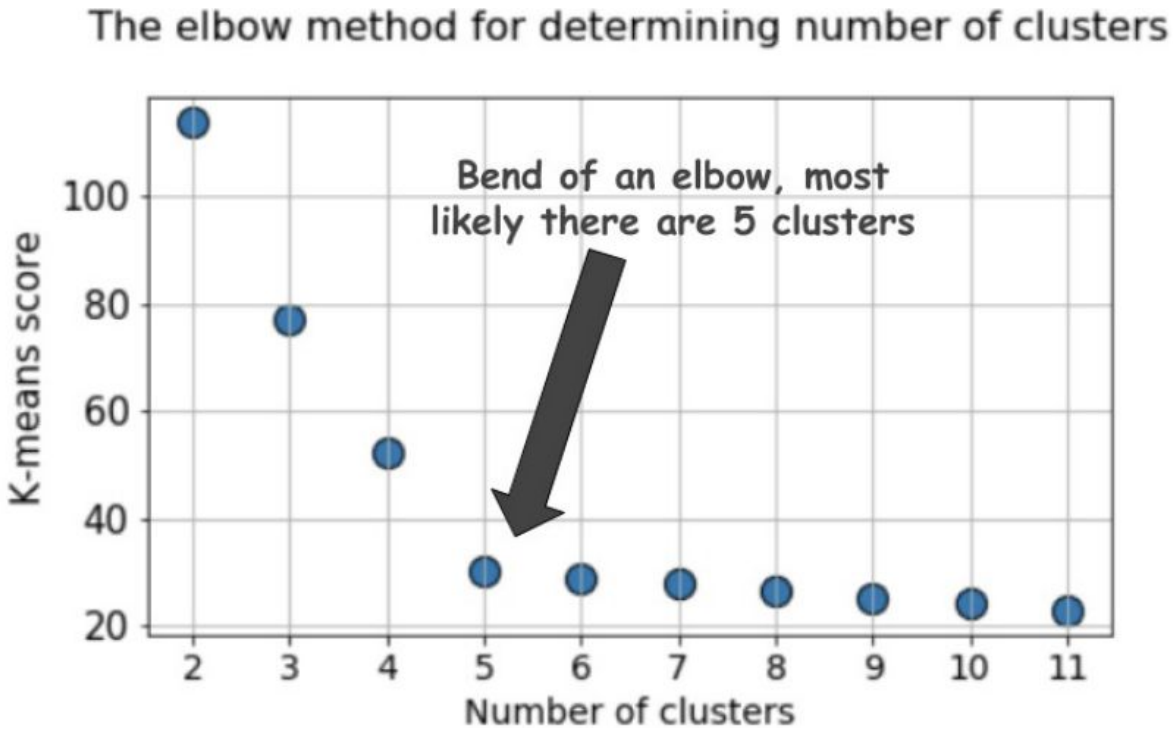
Кластеризация kmeans. Как найти количество кластеров

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Метод Локтя (Elbow Method)

Этот метод помогает определить точку, на которой увеличение числа кластеров перестает значительно улучшать модель.

Идея заключается в том, чтобы найти такое количество кластеров, после которого **уменьшение внутригрупповой дисперсии** (суммы квадратов расстояний от каждой точки к центроиду своего кластера) **становится менее существенным.**



$$R = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2$$

Кластеризация kmeans. Как найти количество кластеров

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Метод силуэта (Silhouette Method) - метод оценки оптимального количества кластеров в алгоритме кластеризации, таком как k-средних.

Метод основан на измерении того, насколько объекты внутри кластера похожи друг на друга и насколько отличаются от объектов в соседних кластерах.

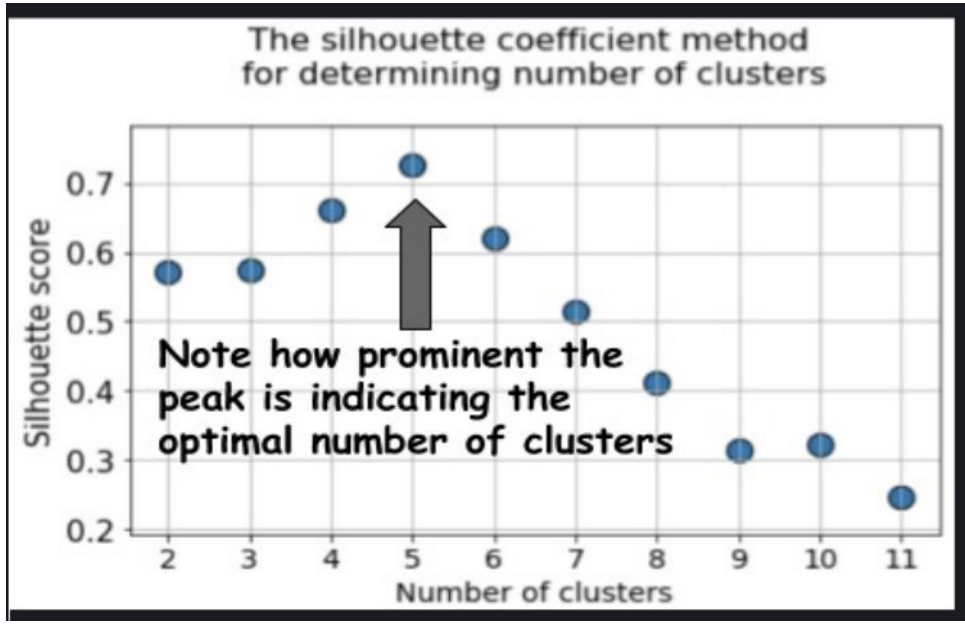


График зависимости $\frac{b-a}{\max(a,b)}$ и кол-ва кластеров

a – среднее внутрикластерное расстояние

b – расстояние до ближайшего кластера

Кластеризация kmeans. Как найти количество кластеров

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Сила силуэта для каждого объекта i в кластере рассчитывается как разница между среднем расстоянием до всех точек внутри **своего** кластера и средней расстоянием до всех точек **ближайшего** другого кластера. Формула для силы силуэта s(i) следующая:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

где:

- a(i) — это среднее расстояние от объекта i до всех других объектов внутри того же кластера (внутрикластерное расстояние).
- b(i) — это среднее расстояние от объекта i до объектов другого кластера, к которому i не принадлежит (межкластерное расстояние).

Кластеризация kmeans. Как найти количество кластеров

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Значение силуэта может варьироваться от -1 до 1:

- s(i)≈1 означает, что объект хорошо соответствует своему кластеру и удален от других кластеров.
- s(i)≈0 означает, что объект находится на границе между двумя кластерами.
- s(i)≈−1 означает, что объект неправильно классифицирован и находится ближе к другому кластеру, чем к своему

Средний силуэт для всех объектов в наборе данных используется как метрика для оценки общей качества кластеризации. Чем выше средний силуэт, тем лучше разделение кластеров.

Кластеризация kmeans. Как найти количество кластеров

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

1. Запустите k-средних с разными значениями k (количество кластеров).
2. Для каждого значения k вычислите силуэт для каждой точки данных:
   a. Вычислите a(i) - среднее расстояние от точки i до всех других точек в том же кластере.
   b. Вычислите b(i) - среднее расстояние от точки i до всех точек ближайшего кластера (кластера, к которому точка не принадлежит).
   c. рассчитайте силуэт для точки i
3. Для каждого значения k вычислите средний силуэт для всех точек данных в этом кластере.
4. Выберите значение k, при котором средний силуэт максимален.

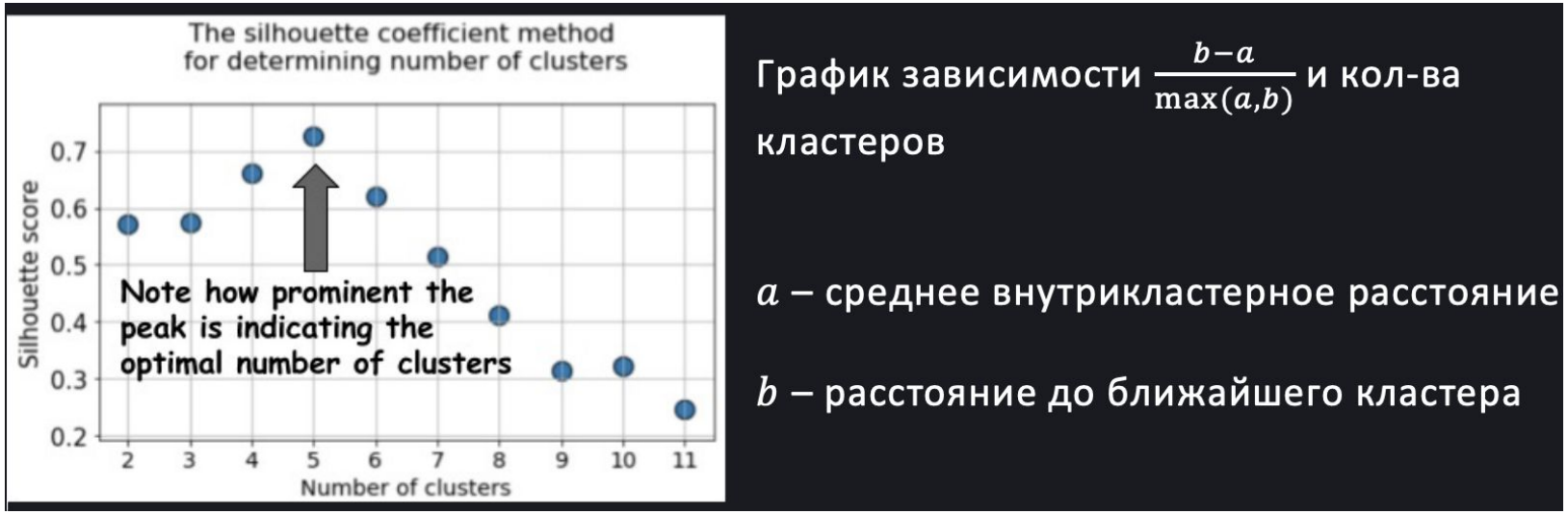


График зависимости $\frac{b-a}{\max(a,b)}$ и кол-ва кластеров

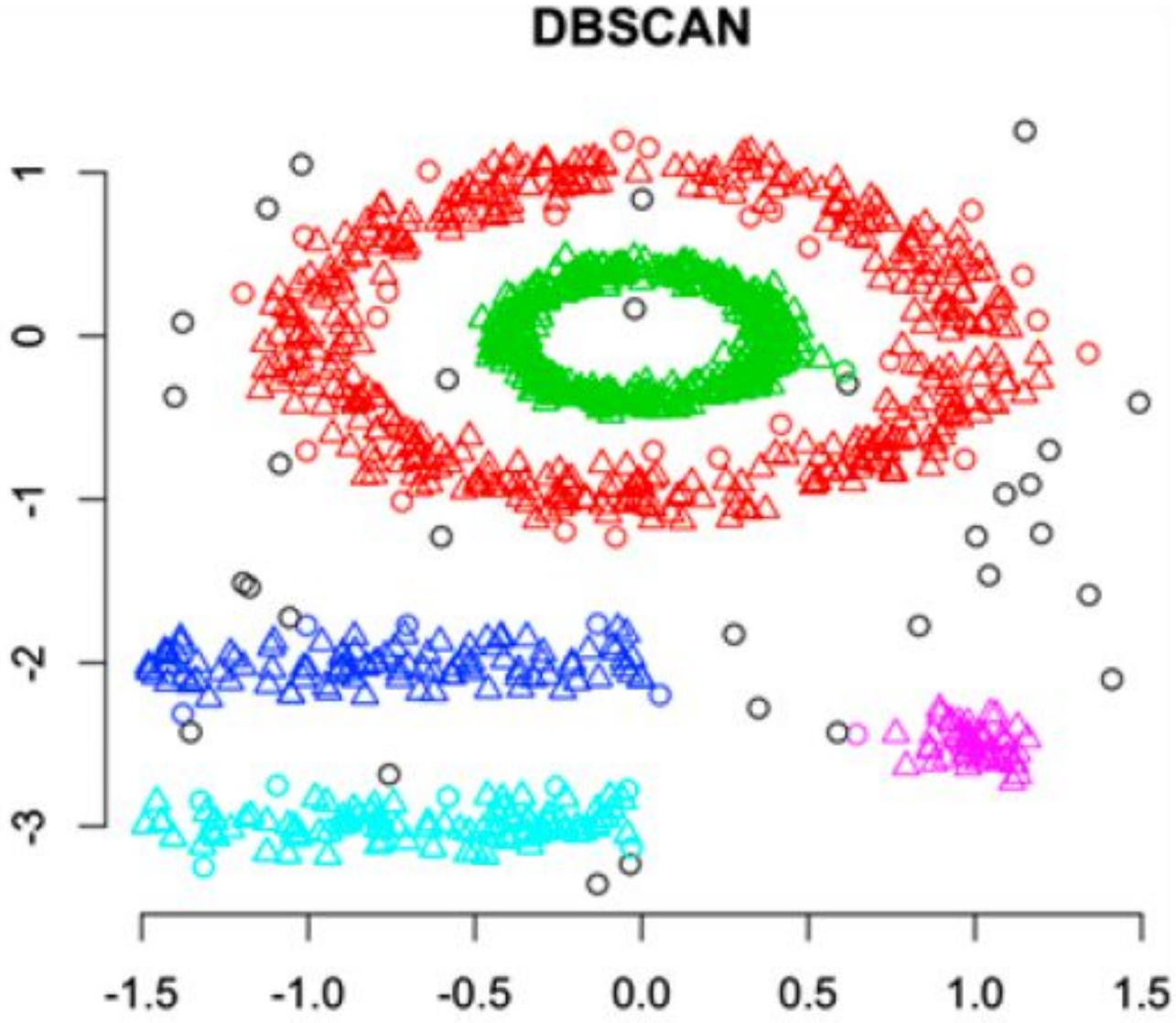
a – среднее внутрикластерное расстояние

b – расстояние до ближайшего кластера

Методы кластеризации.DBSCAN

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Алгоритм DBSCAN (Density-based spatial clustering of applications with noise) развивает идею кластеризации с помощью выделения связных компонент.

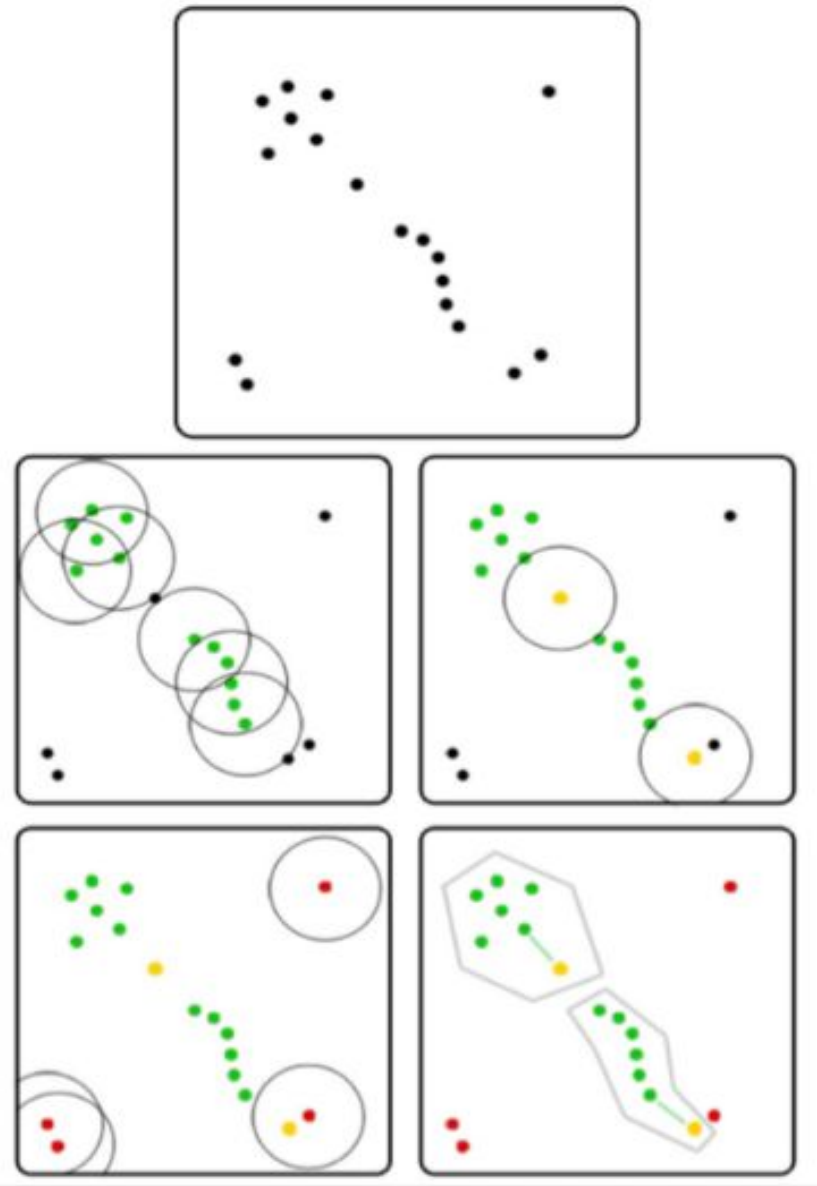


Методы кластеризации.DBSCAN

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Метод зависит от двух параметров: радиуса эпсилон и минимального числа точек в окрестности k

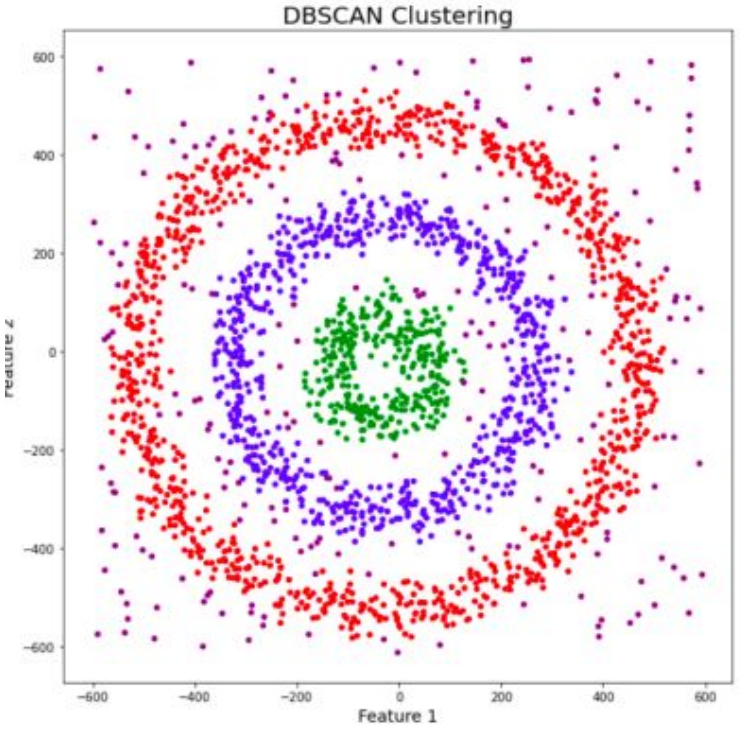
- Рассматриваем объекты как ядра, вокруг которых собираются другие объекты. Точка является ядром, если в ее эпсилон окрестности k точек
- Если точка находится в эпсилон-окрестности точки-ядра, но сама по себе не является точкой-ядром, то она считается точкой-граничной и также присоединяется к ближайшему кластеру.
- Если ядра связаны, то они и достижимые из них объекты образуют кластер
- Если точка не входит в эпсилон-окрестность ни одной другой точки-ядра, то она считается выбросом или "шумом". Точки-шум не присоединяются ни к одному кластеру



Методы кластеризации.DBSCAN

DBSCAN сам определяет количество кластеров. Кластеры могут иметь вид протяжённых лент или быть вложенными друг в друга как концентрические гиперсферы.

DBSCAN — один из самых сильных алгоритмов кластеризации, но работает он, как правило, заметно дольше, чем mini-batch K-means, к тому же весьма чувствителен к размерности пространства признаков, поэтому используется на практике DBSCAN только тогда, когда успевает отрабатывать за приемлемое время.



доп материалы

- https://education.yandex.ru/handbook/ml/article/klasterizaciya

Что делать после того, как наконец выброс определили?

1. **Удаление выбросов**

Применимо если:

- Выбросов немного (например, <5–10%)
- Данные после удаления остаются репрезентативными

2. Замена выбросов

Варианты:

- **На медиану или среднее** по колонке:
- **На соседние значения** (если данные временные → можно взять соседние по времени)
- **На предсказание модели** (например, обучить регрессию на нормальных данных и предсказать выбросам)

3. Ограничение значений (Winsorization)

- Замена слишком больших/маленьких значений на граничные.

Что делать после того, как наконец выброс определили?

4. Использование устойчивых моделей (robust models)

Если ты **не удалять выбросы**, можно использовать алгоритмы, **устойчивые к выбросам**:

- RandomForest, GradientBoosting, XGBoost — хорошо справляются с шумами
- RobustScaler вместо StandardScaler при нормализации
- Регрессия с регуляризацией (HuberRegressor, RANSAC)

5. Оставить выбросы как полезную информацию

Иногда выбросы — это не ошибка, а **важные сигналы**:

- **Мошенничество**, редкие **болезни**, уникальные **VIP-клиенты**
- В таких случаях выбросы не удаляют, а **используют как целевой класс**

Что делать после того, как наконец выброс определили?

Общее правило:

Выбросы удаляют или обрабатывают только на тренировочной выборке.
Тестовая выборка **не трогается** — она моделирует "реальные новые данные".

Почему так?

- Тест используется **только для оценки** финальной модели.
- Если мы чистим тестовые данные, мы **искусственно улучшаем метрики**.
- В реальности модель будет получать и "грязные" данные, включая выбросы — её задача научиться с ними справляться.

Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Спасибо за внимание