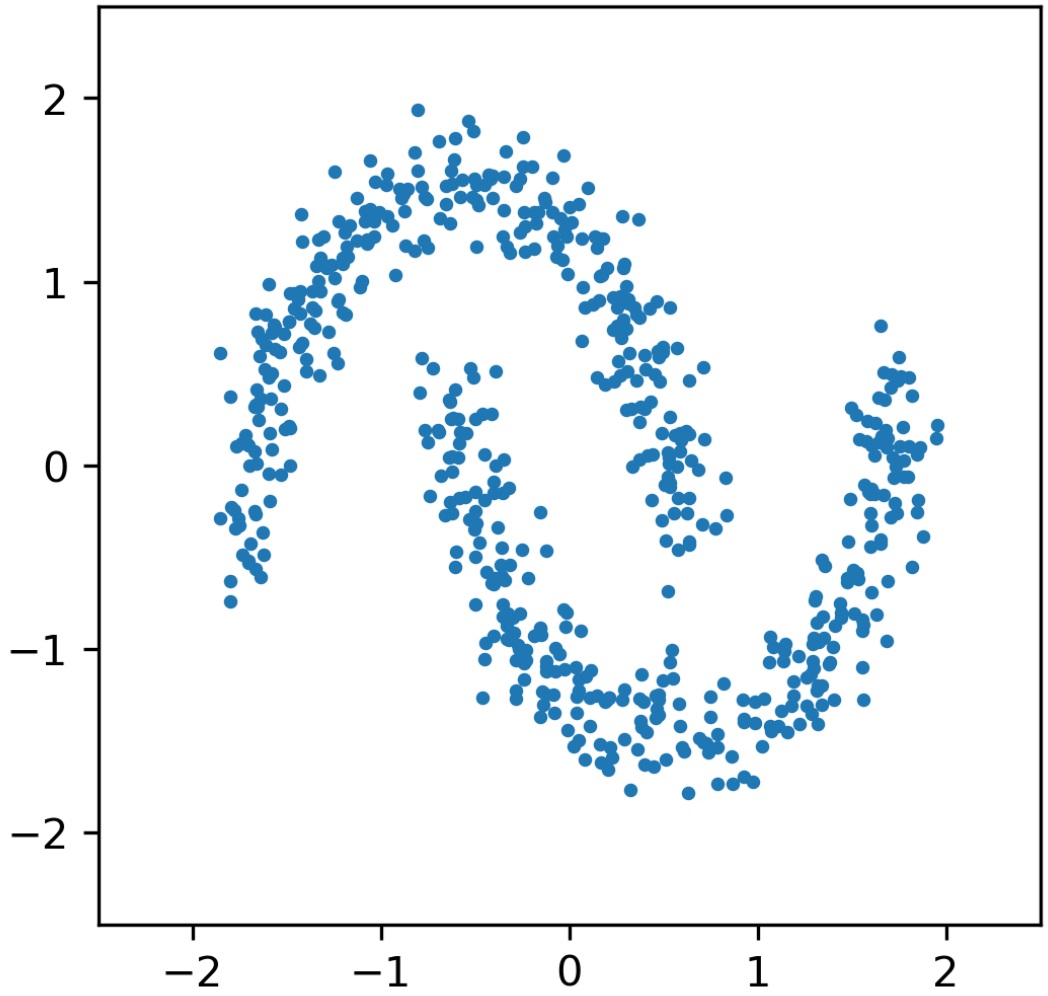
Машинное обучение в гидрометеорологии

М.А. Криницкий

ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

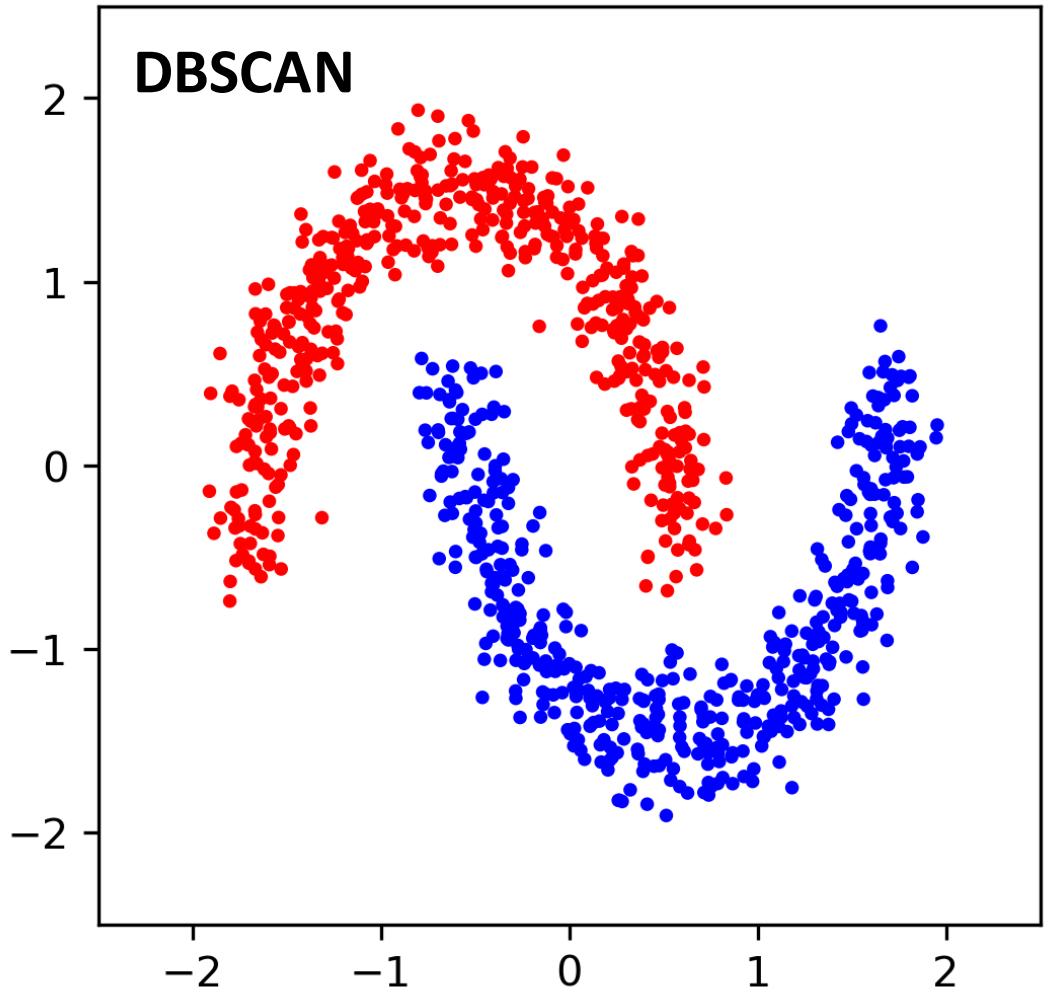
- «Обучение без учителя»
  - поиск структуры в данных



ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

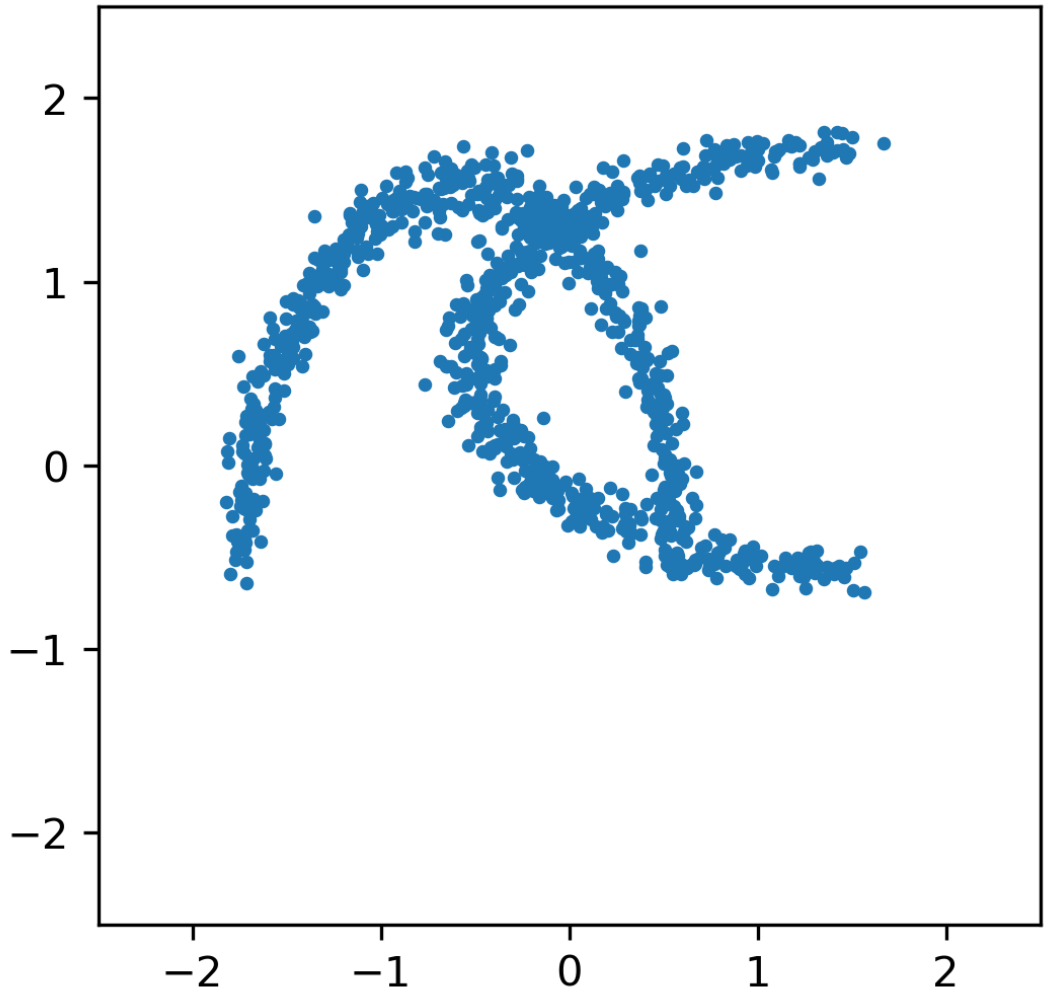
- «Обучение без учителя»
  - кластеризация



ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
  - кластеризация

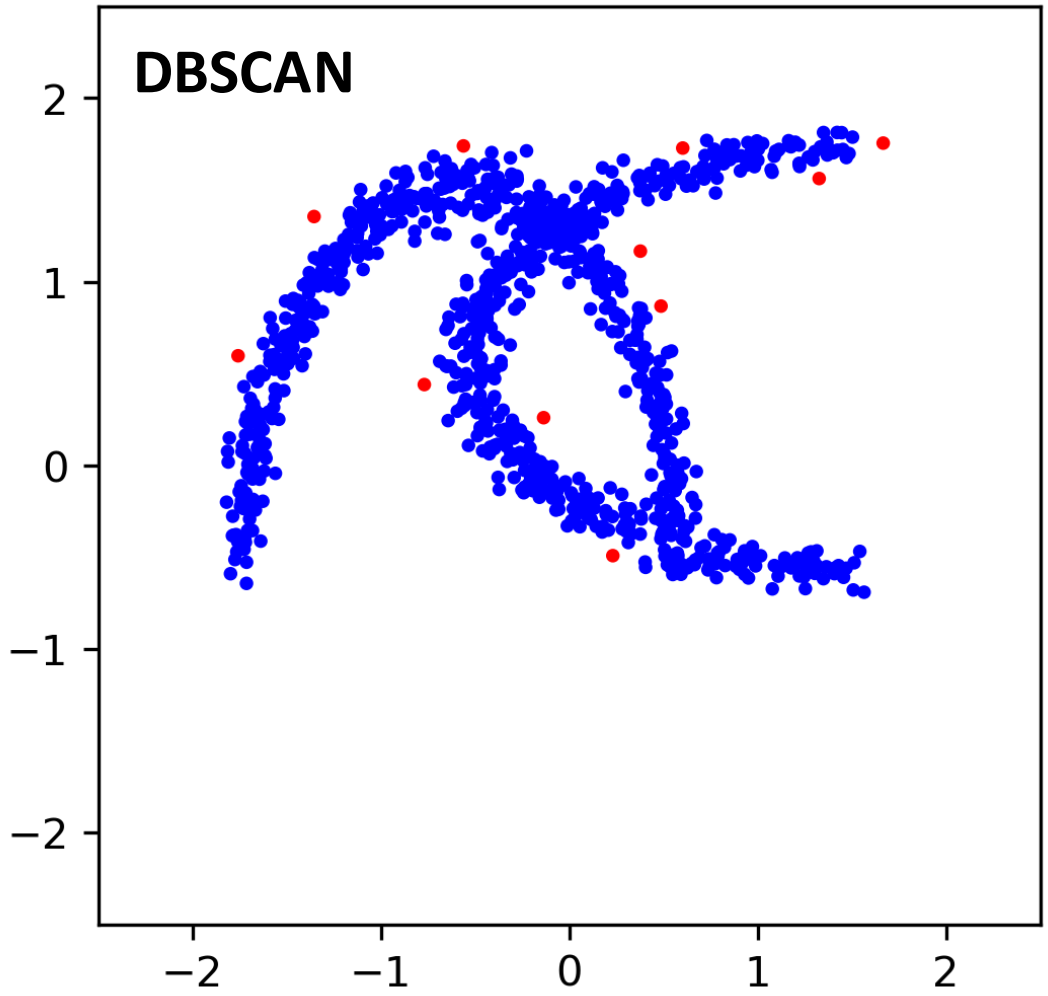


ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
  - кластеризация

Всегда ли есть решение,
которое мне понравится?

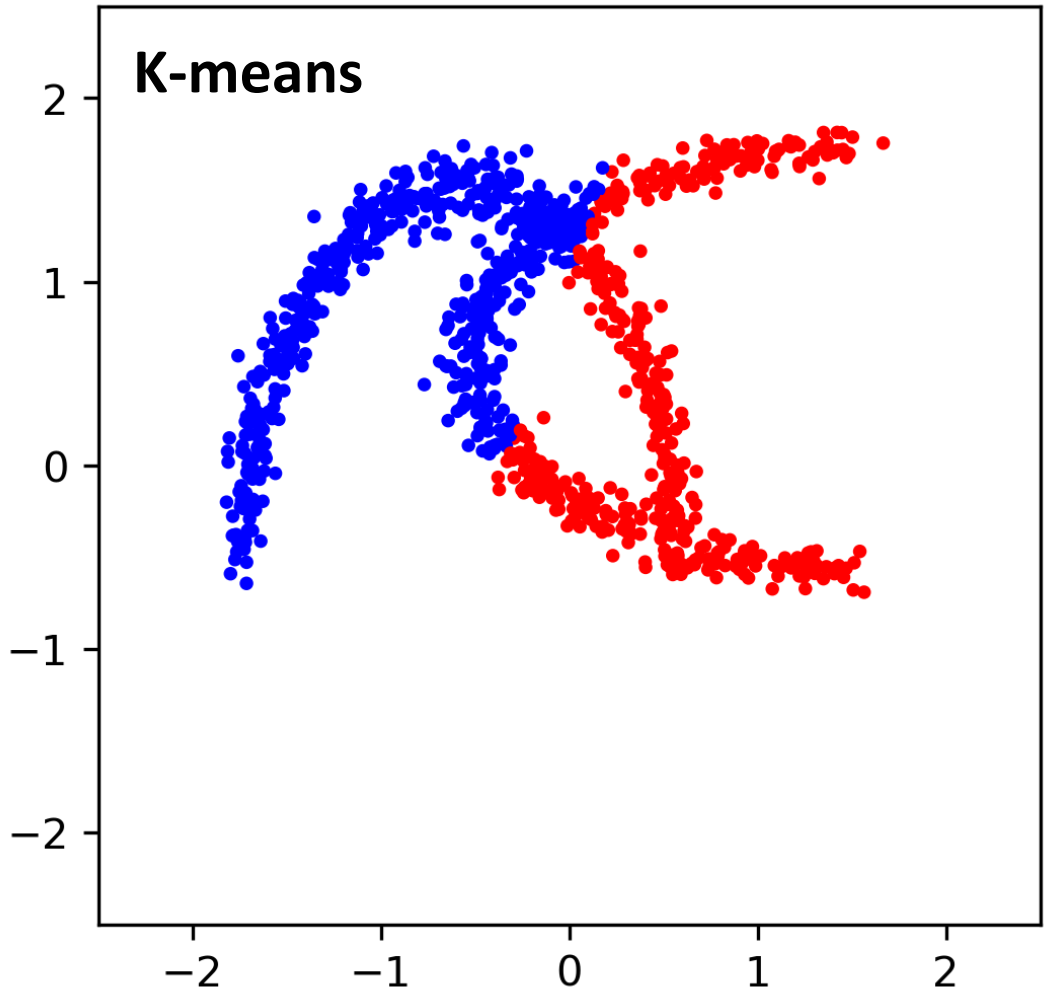


ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
  - кластеризация

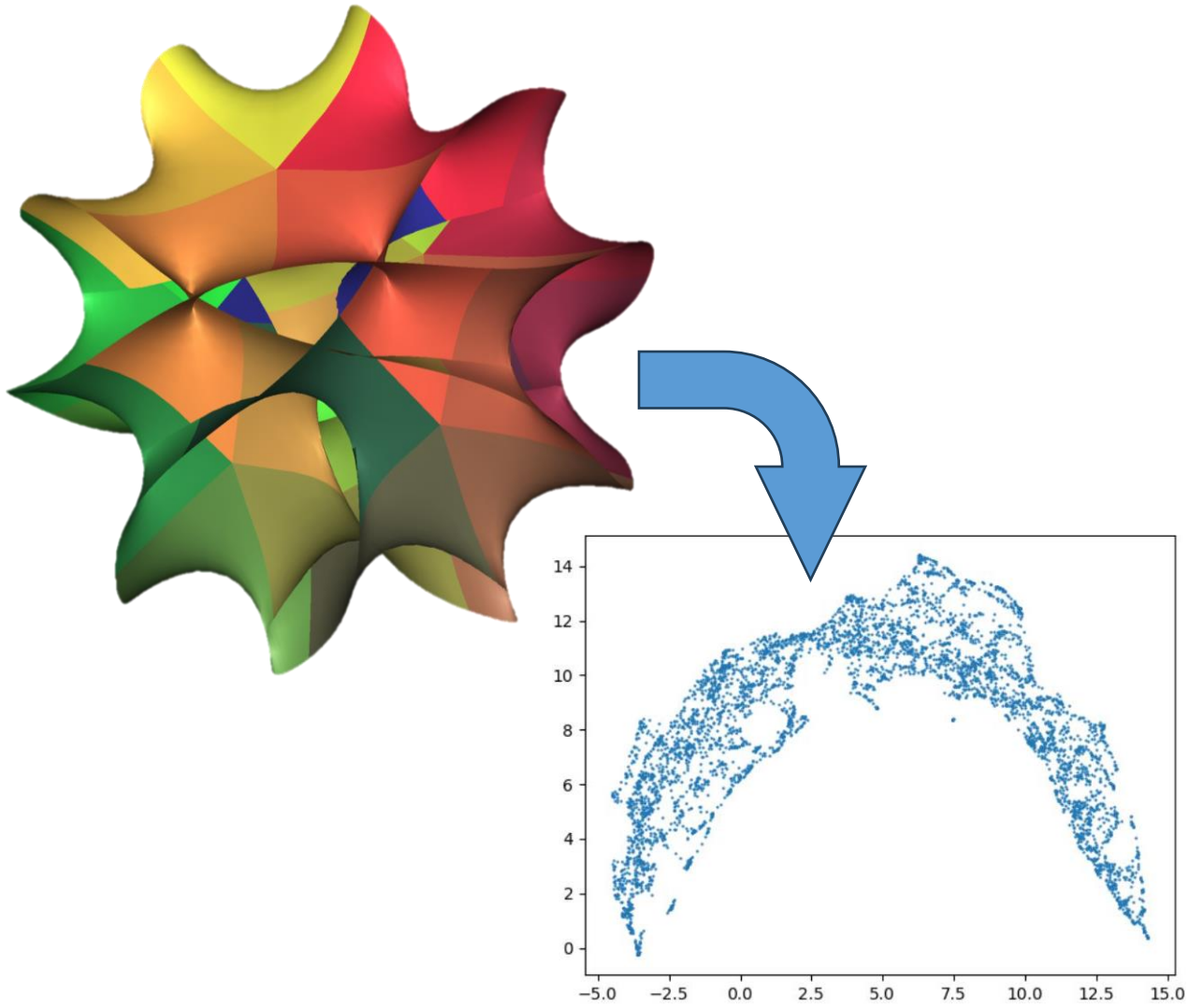
Всегда ли есть решение,
которое мне понравится?



ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
  - снижение размерности



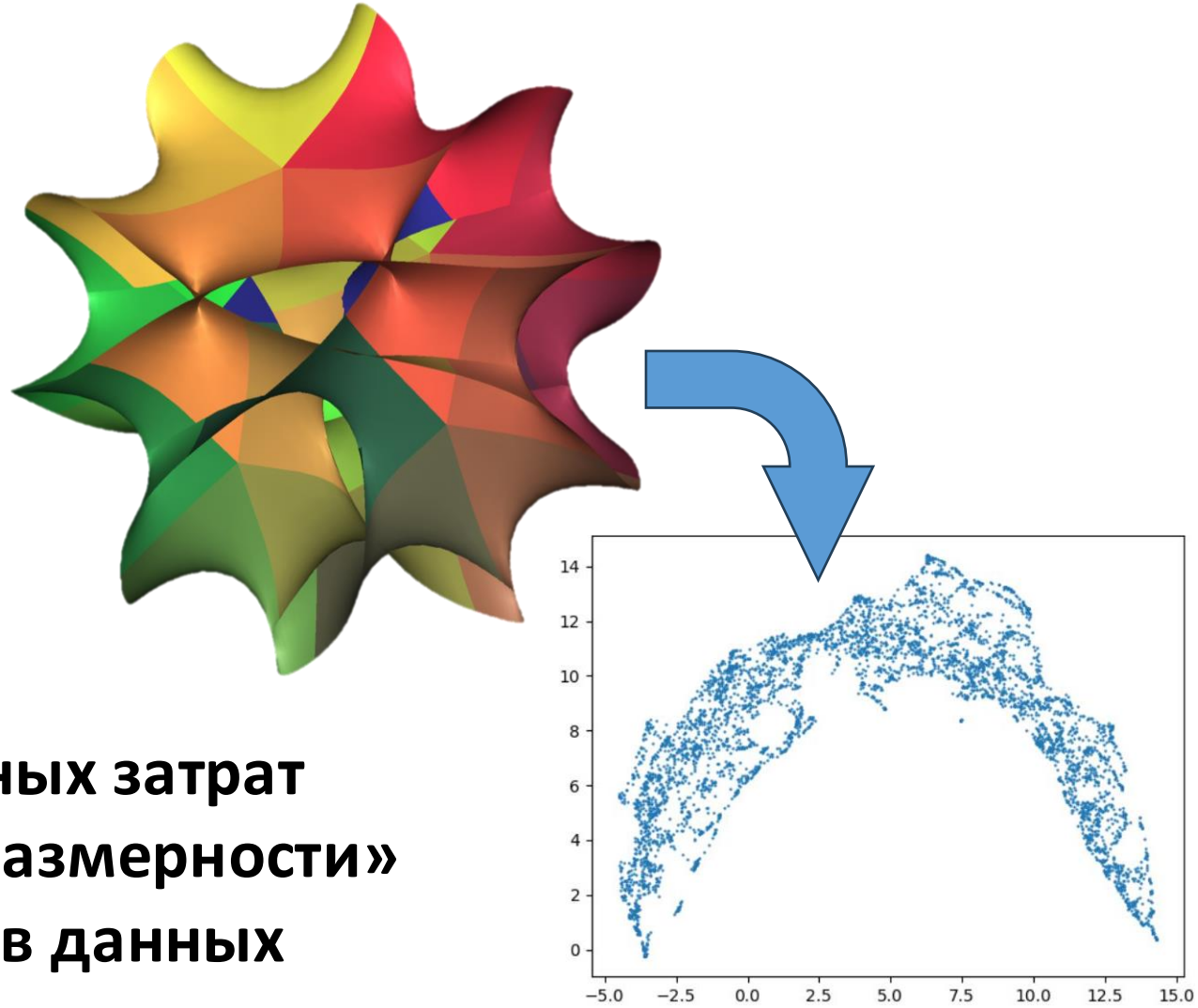
ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
  - снижение размерности

что я хочу?

признаковое описание сниженной размерности



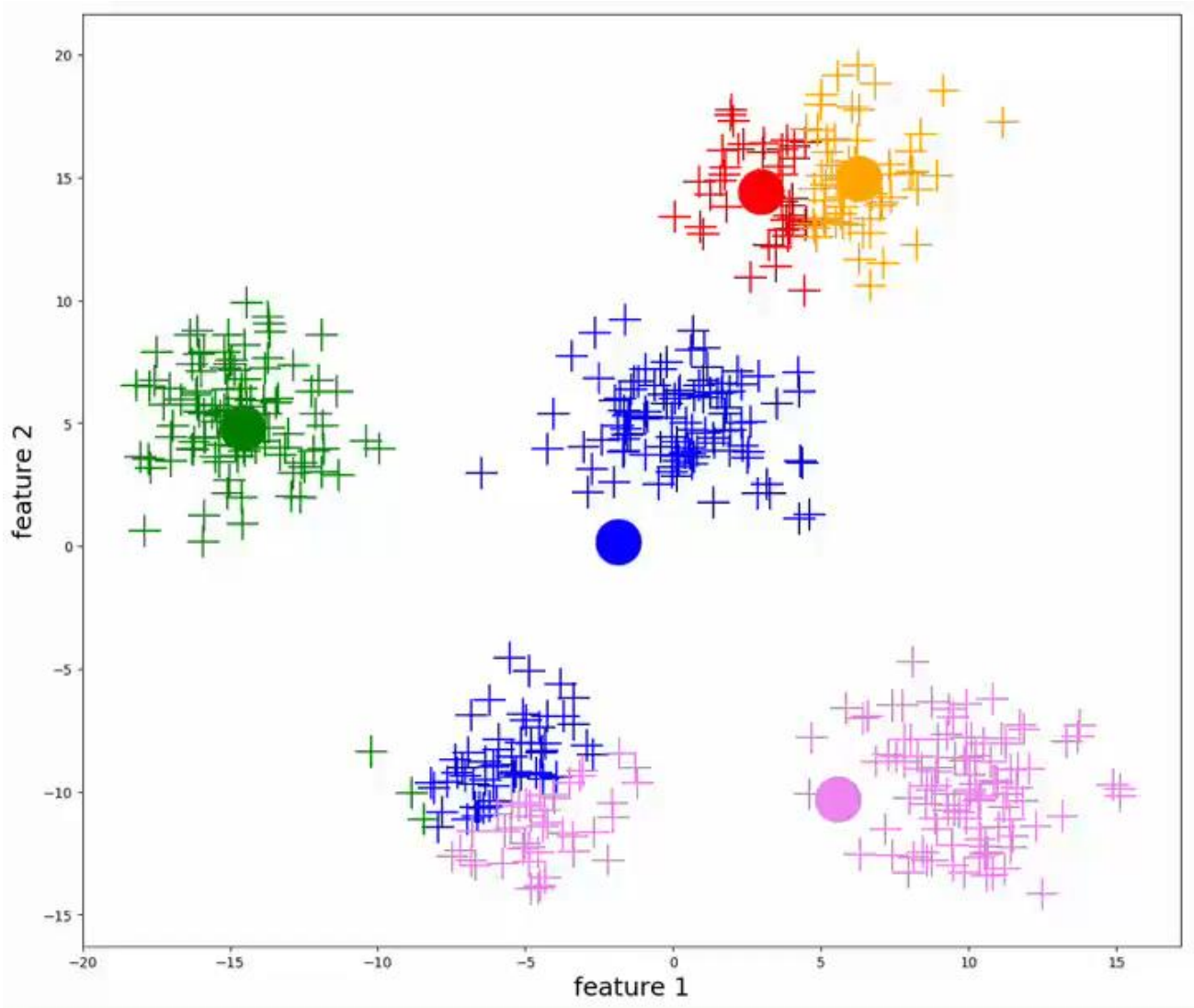
зачем?

- **визуализация данных**
- **снижение вычислительных затрат**
- **борьба с «проклятием размерности»**
- **снижение уровня шума в данных**

K-Means clustering

1. Инициализация:
   a. количество кластеров K;
   b. начальное положение центроидов кластеров
   c. критерий останова процедуры кластеризации
2. На каждой итерации:
   a. объекты приписываются к ближайшему центроиду
   b. положение центроидов пересчитывается как новый центр масс каждого кластера
   c. проверяется критерий останова

K-Means clustering



K-Means clustering

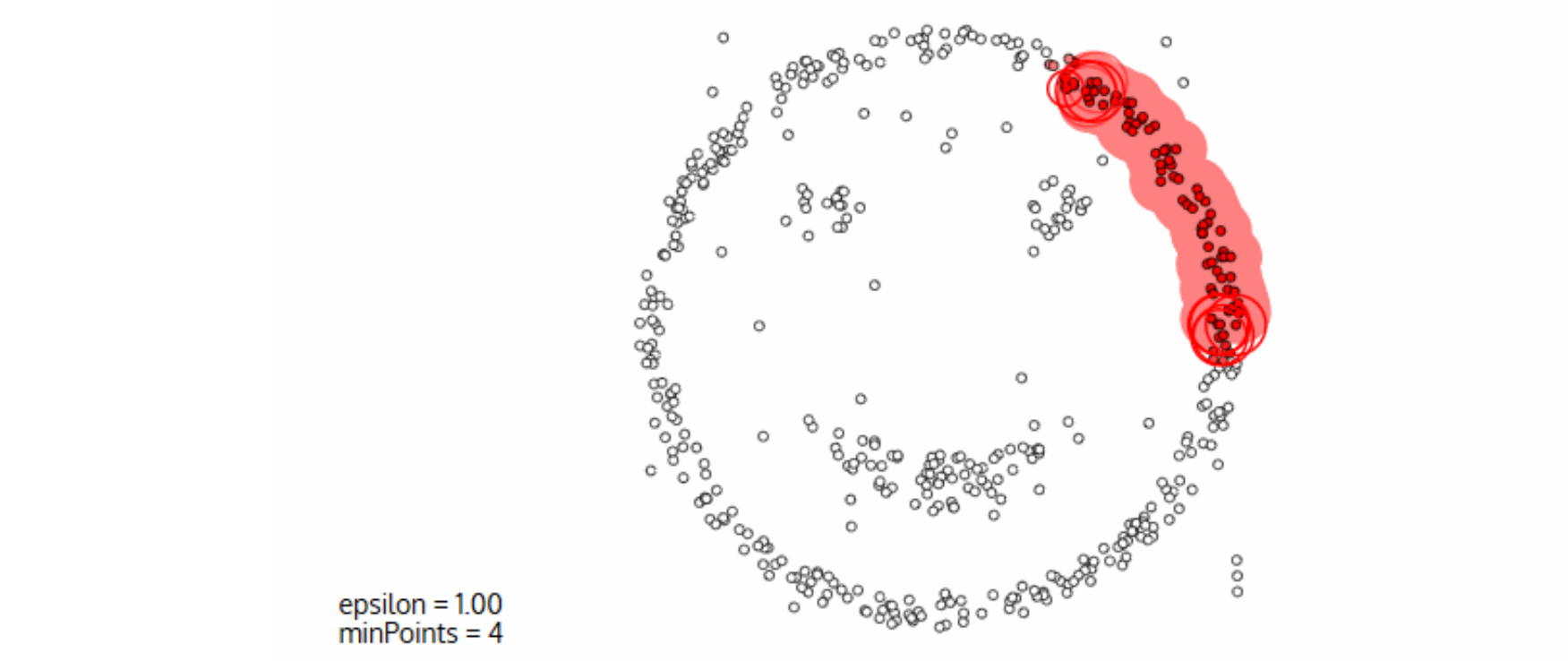
- Очень быстрый метод
- Нужно задавать количество кластеров K (гиперпараметр)
- Используется евклидово расстояние в качестве меры дистанции
- Результат зависит от начальной инициализации
  - нет гарантии воспроизводимости результата
- Нет встроенной возможности присовокупить новые объекты в результат кластеризации
  - нет гарантии, что в обновленной коллекции данных «старые» объекты разобьются по кластерам так же, как в предыдущей коллекции

DBSCAN

Density-based Spatial Clutering of Applications with Noise

1. Инициализация
   a. Радиус окрестности ϵ
   b. минимальное кол-во m объектов кластера
2. На каждой итерации:
   a. выбирается объект, не отнесенный ни к одному кластеру
      i. для такого объекта отбираются объекты в окрестности ϵ
         - если таких объектов меньше m, объект считается выбросом, процесс продолжается с п. (a)
      ii. иначе кластеризация продолжается в следующем порядке:
         - создается метка кластера, объект и его соседи приписываются к этому кластеру
         - для каждого объекта-соседа повторяется процедура поиска соседей в окрестности ϵ, они приписываются к этому же кластеру.
3. Процедура продолжается до полного исчерпания всех объектов коллекции

DBSCAN

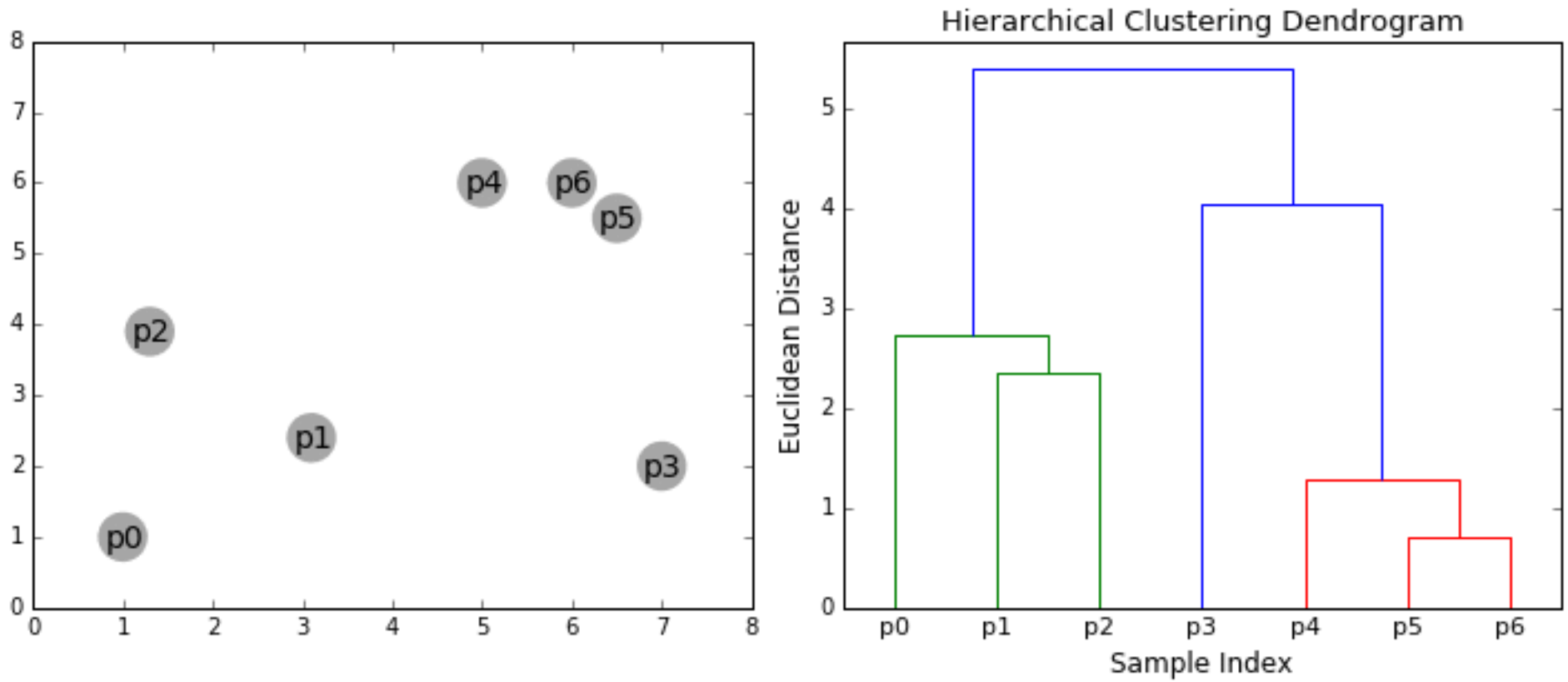


DBSCAN

Density-based Spatial Clutering of Applications with Noise

- Гиперпараметры: ϵ, m, метрика (мера дистанции)
- НЕ требуется предполагать количество кластеров
- Встроенная функциональность обработки выбросов
- Обрабатывает кластеры произвольных форм и размеров лучше K-Means
- Медленнее, чем K-Means
- Эффективность снижается при повышении различий в плотности объектов в кластерах

Агломеративная иерархическая кластеризация



Агломеративная иерархическая кластеризация

1. Инициализация
   - Выбирается мера дистанции между кластерами (напр., средняя попарная дистанция между объектами кластеров)
   - Каждый объект – отдельный кластер
2. Итерационно:
   a. вычисляются попарные дистанции между кластерами
   b. два кластера на минимальной дистанции объединяются в один новый
3. Процедура продолжается до объединения всех кластеров в один
4. Результат кластеризации выбирается по мере качества кластеризации (все результаты промежуточных кластеризаций доступны для выбора)

Агломеративная иерархическая кластеризация

- Нет необходимости заранее задавать кол-во кластеров
- Алгоритм нечувствителен к выбору меры дистанции между объектами (лишь она была метрикой)
- Выбор подходящего результата кластеризации производится на основании меры качества

Качество кластеризации

- Не с чем сравнивать => нет «истины»
- Нет «самого правильного» результата кластеризации – есть много вариантов в зависимости от метода и гиперпараметров
- На основании качества кластеризации нужно выбрать метод и гиперпараметры

Качество кластеризации

- Inertia – мера компактности кластеров. Специфична для отдельных кластеров.

$$I_c = \sum_C (x_i - C_c)^2$$

- Чем ниже, тем лучше (кластер более компактен)
- Неотрицательный
- Нет верхней границы
- Неадекватно характеризует качество для «растянутых» кластеров
- автоматически вычисляется в K-means (в sklearn)

Качество кластеризации

- Silhouette Score – мера компактности кластеров и одновременно их разделимости

для объекта x_i кластера C:

$a_i = \frac{1}{|C|}\sum d_{ij}$ - среднее внутрикластерное расстояние (i,j – в кластере C)

$b_i = \frac{1}{|C_J|}\sum d_{ij}$ - среднее расстояние до «ближайшего» кластера C_J (i,j - в разных кластерах)

$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$ - коэффициент силуэта для объекта x_i

- от -1 до 1
- Чем выше, тем лучше
- S_C - средний по кластеру коэффициент силуэта
- Меры качества кластеризации: максимальный S_C, средний S_C

Идентификация аномалий

- Цель: идентификация объектов/событий, аномальных для коллекции данных
  - Применения: обнаружение ошибок, помех, аномального поведения или особых состояний систем
  - Нет четкого определения
  - Все соображения базируются на идее о ненормальности объектов/событий, так или иначе не укладывающихся в ОСНОВНУЮ закономерность данных в коллекции
  - часто предполагается, что аномальных объектов в коллекции МАЛО
  - => так или иначе имеет смысл (а) моделировать основную закономерность и (б) оценивать меру «выпадения» объекта из основной закономерности

Идентификация аномалий

- DBSCAN
- Аномалии в контексте моделирования распределения данных

Идентификация аномалий

- Isolation Forests
  - Строятся на Isolation Trees (iTree)
    - iTree: строятся как Decision Trees за исключением отсутствия оптимизации: признак и правило принятия решения выбирается произвольно;
    - iTree строится до состояния 1 объекта на лист или до максимально допустимой глубины;
    - Каждое iTree строится на подмножестве коллекции данных;
    - Чем более аномальный объект, тем быстрее (на более ранних ветвлениях) он будет «отсечен» в отдельный лист при равномерном сэмплировании признака и правила разбиения;

Идентификация аномалий

- Isolation Forests
  - Для каждого объекта x_i измеряется длина пути h_{ij} до «его» листа в каждом iTree$_j$
  - Длины путей для каждого объекта осредняются по всем iTree: $E_i = \mathbb{E}_j h_{ij}$
  - Чем меньше длина пути, тем более вероятно, что объект аномален

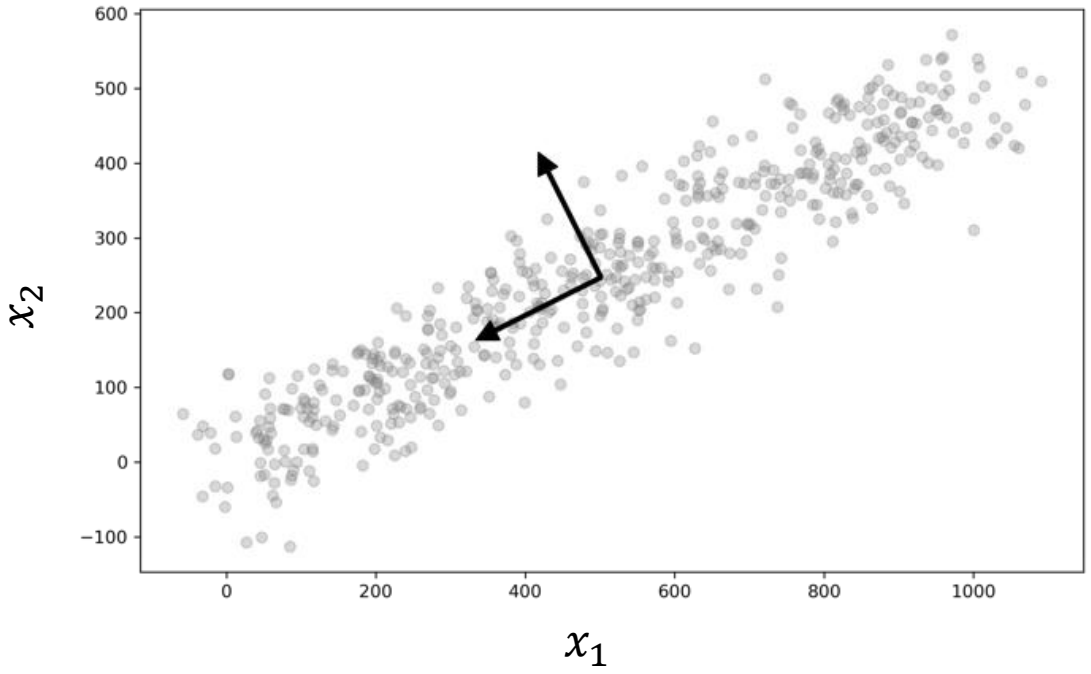
$$s_i = 2^{-\frac{E_i}{c_m}}$$

  - s_i близкое к 1 – высока вероятность аномалии;
  - s_i менее 0.5 – аномалия маловероятна

$$c(m) = \begin{cases} 2H(m-1) - \frac{2(m-1)}{n} & \text{for } m > 2 \\ 1 & \text{for } m = 2 \\ 0 & \text{otherwise} \end{cases}$$

Снижение размерности: PCA

- Метод главных компонент (PCA, principal components analysis)



1. Стандартизировать признаки: $\tilde{X}$
2. Рассчитать матрицу ковариаций (стандартизированных) признаков $\tilde{X}$
3. Найти собственные векторы и собственные значения матрицы ковариаций (стандартизированных) признаков
    1. собственные векторы – направления максимальной дисперсии данных
    2. собственные значения – величина объясненной дисперсии данных для компоненты
4. Отсортировать собственные векторы по убыванию собственных значений
5. Отобрать top-k векторов и значений, в сумме объясняющих заданную долю дисперсии в данных, составить матрицу главных компонент W
6. Спроецировать данные на векторы top-k главных компонент: $Z = XW$

https://setosa.io/ev/principal-component-analysis/

Снижение размерности: UMAP

Uniform Manifold Approximation and Projection (UMAP)

- Строит граф данных (каждый объект коллекции – узел графа) на основании близости.
  - К каждому объекту находятся К ближайших соседей согласно метрике в пространстве оригинальных признаков;
  - граф составляется из ребер, которым присваиваются веса согласно дистанциям: $w_{ij} = \exp(-\frac{d_{ij}}{\sigma_i})$
- Строит граф в низкоразмерном пространстве (2D, 3D), по структуре соответствующий графу в высокоразмерном пространстве признаков; узлы произвольно расположены в пространстве;
- Расположение точек в низкоразмерном пространстве оптимизируется:
  - минимизируется отклонение весов в высокоразмерном и низкоразмерном пространствах)
  - процедура оптимизации – итерационная
  - учитываются «силы притяжения» для объектов, расположенных близко и «силы отталкивания» для удаленных объектов

https://pair-code.github.io/understanding-umap/

Обучение без учителя

типы задач:

- ...
- «Обучение без учителя»
  - Кластеризация
  - Снижение размерности
  - Идентификация аномалий
  - Аппроксимация распределения данных
- ...