Занятие № 10

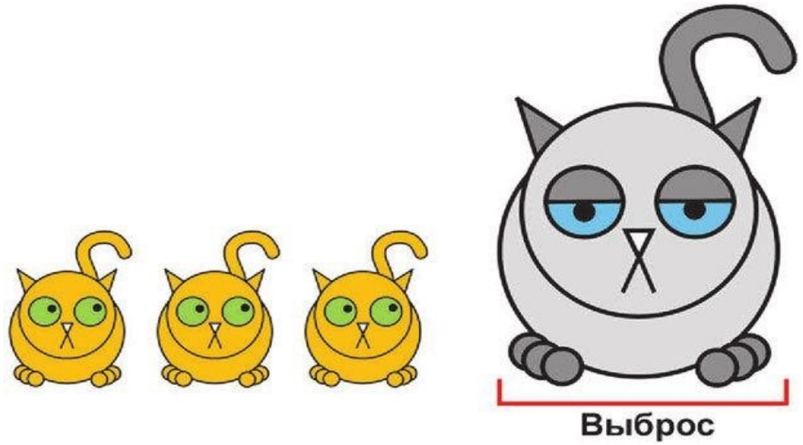
Поиск выбросов и генерация новых признаков

Содержание

1. Что такое выброс
2. Как их отлавливать?
3. Практика.

Что такое выброс?

Выбросы - точки данных, которые не пренадлежат определенной популяции



Причины:

- ошибок в данных
- наличия шумовых объектов
- присутствия объектов «других» выборок
- наличие причины приводящей к выбросы, но слабо проявленной в выборке.

Что такое выброс?

Выбросы так же можно поделить
По мерности:

1) Одномерные - выбросы по одному признаку
2) Многомерные - выброс по подмножеству признаков

По типу:

1) Точечные выбросы - единичные точки, выбивающиеся из общей картины.
2) Контекстуальные выбросы – значения которые являются выбросами только в определенном контексте.
3) Коллективные выбросы – отклонения от нормальных значений не единичной точки, а целой группы

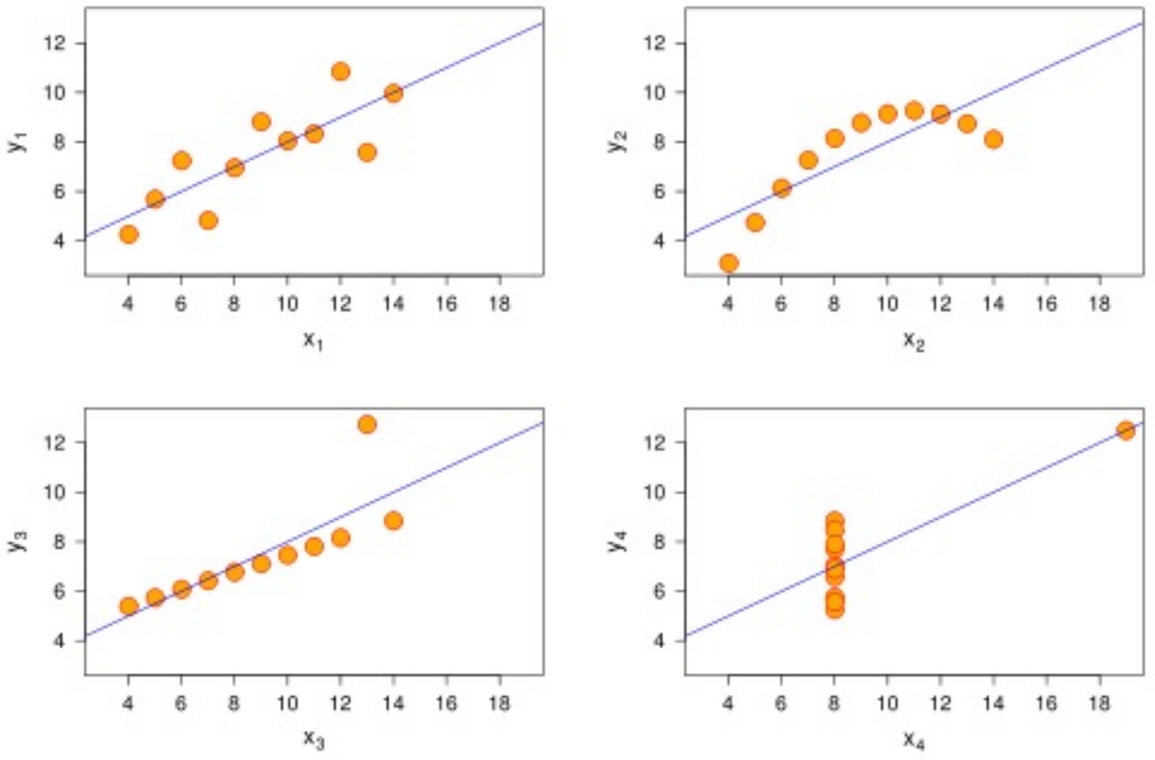
Почему с ними нужно бороться?

1) Многие алгоритмы машинного обучения чувствительны к разбросу и распределению значений признаков за счет механизма оптимизации с функциями потерь имеющих экстремальные значения на больших отклонениях.

2) Выбросы во входных данных могут исказить статистики или результатов статистических расчетов

3) Введение в заблуждение исследователя и искажение выводов о данных

Квартет Энскомба



Как их отлавливать?

- **Статистические тесты**
  - на основе стандартного отклонения (2σ - 3σ)
  - на основе межквартильного расстояния $[(x_{25} - 1{,}5 \cdot (x_{75} - x_{25})),\ (x_{75} + 1{,}5 \cdot (x_{75} - x_{25}))]$
  - *Q*- тест Диксона
  - Критерий Граббса
  - и др.

- **Модельные тесты**
  - построение модель, которая описывает данные по исследуемому признаку (выбросы плохо описываеются моделью)

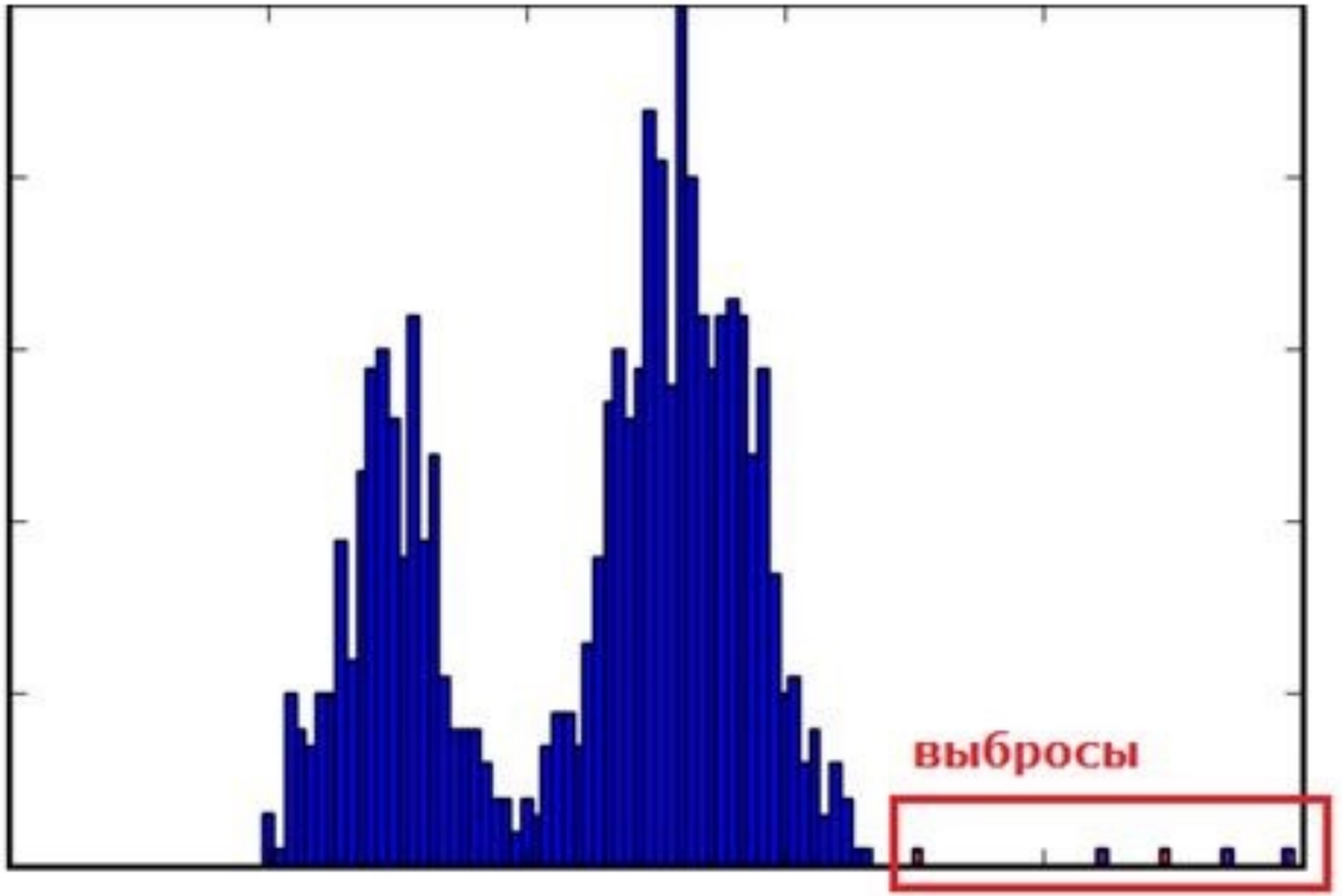
Как их отлавливать?

- **Метрические методы**
  KNN (LOF, Расстояние Махаланобиса)
  DBScan

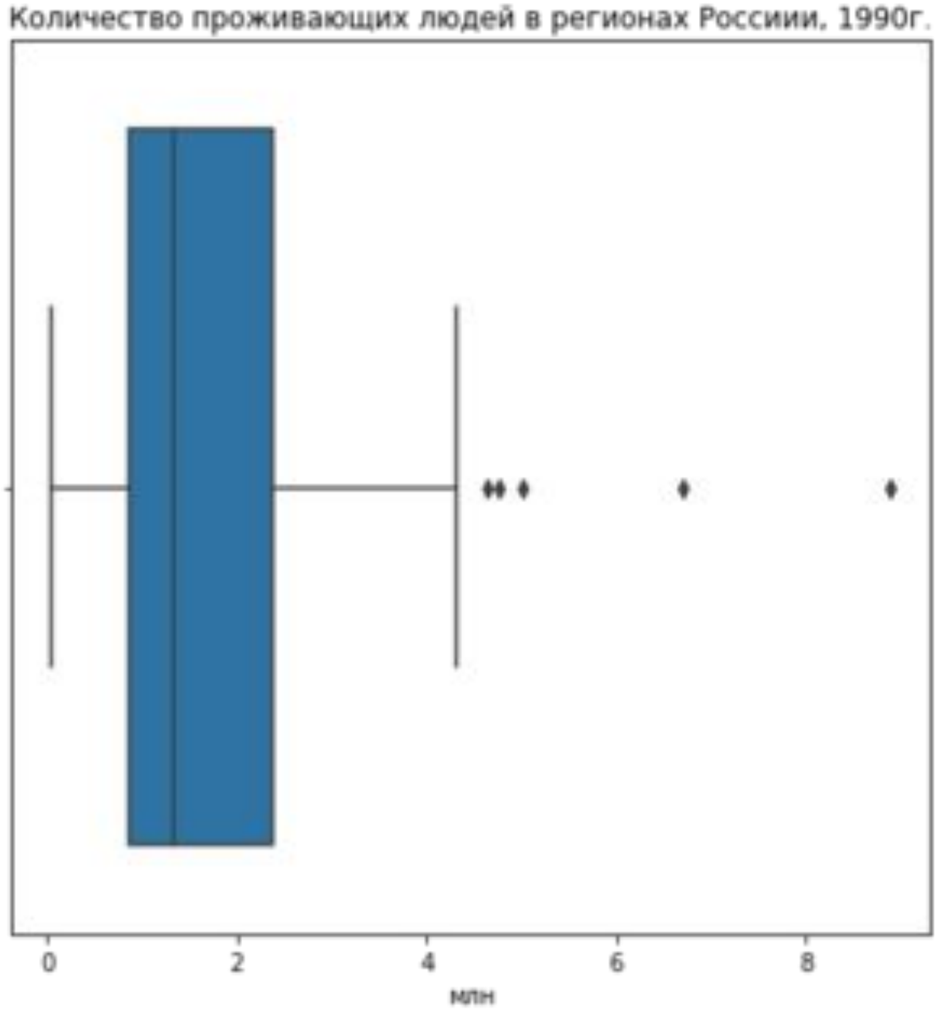
- **Методы машинного обучения**
  Метод опорных векторов для одного класса σ
  Изолирующий лес

- **Итерационные методы**
  построение оболочек в n-пространстве

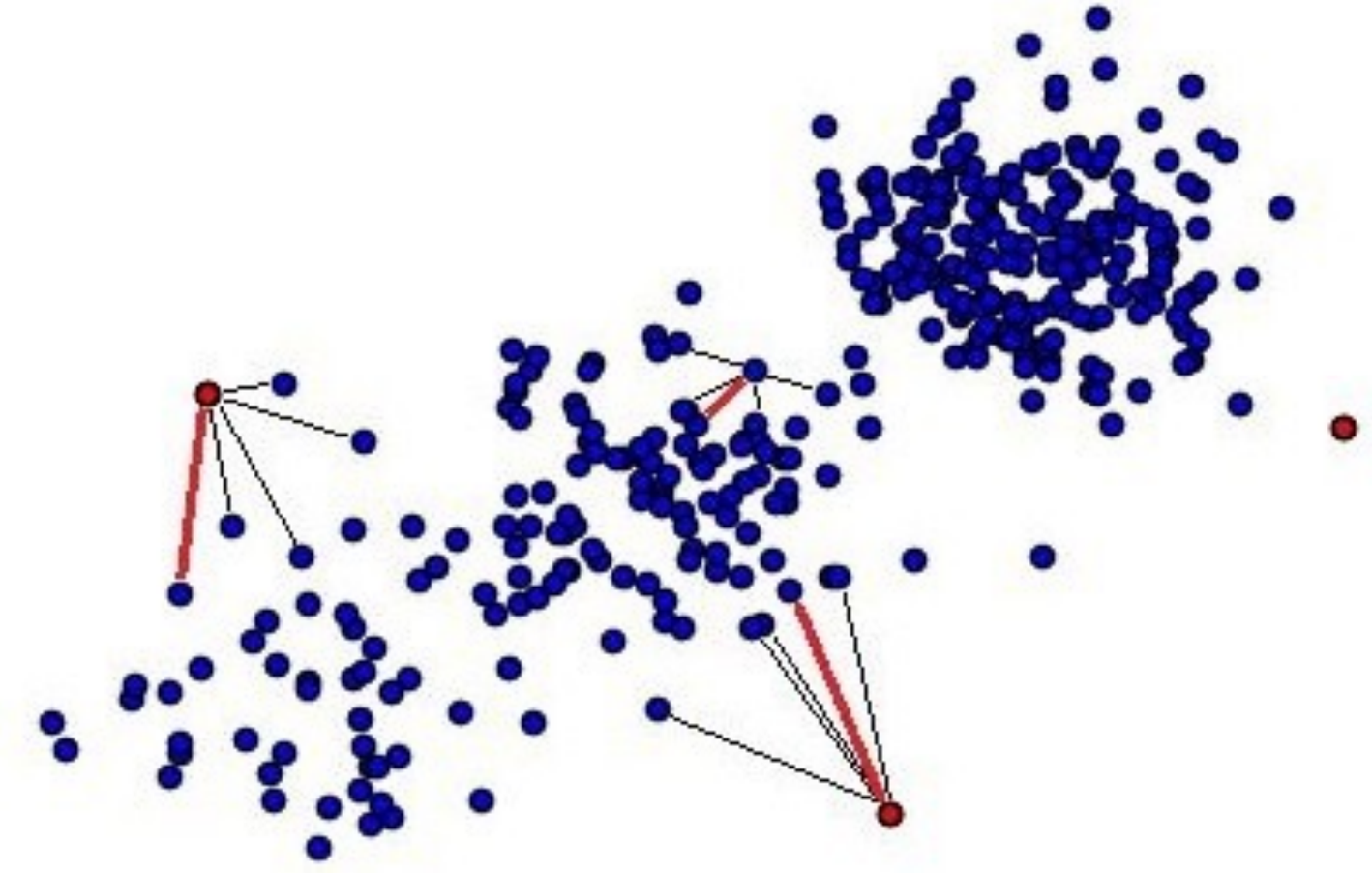
Как их отлавливать?
Выбросы из выборки



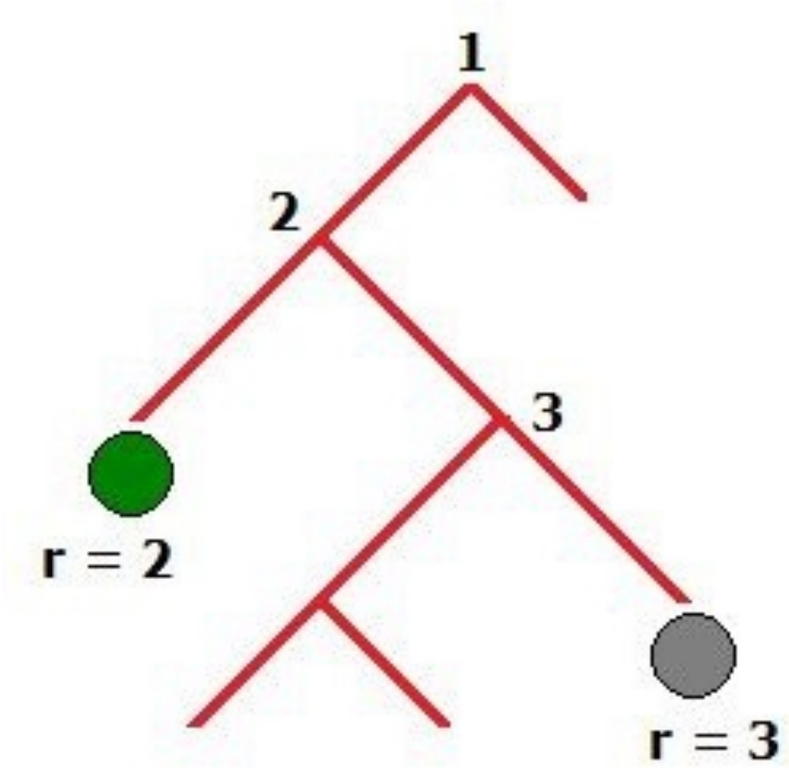
Как их отлавливать?
Выбросы из выборки



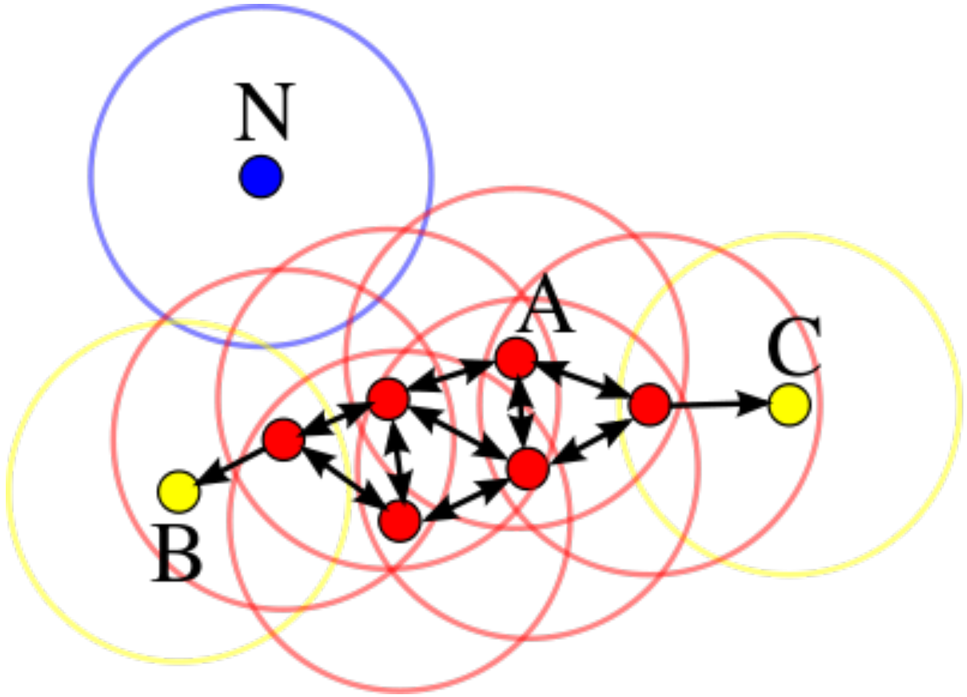
Как их отлавливать?
KNN (LOF)



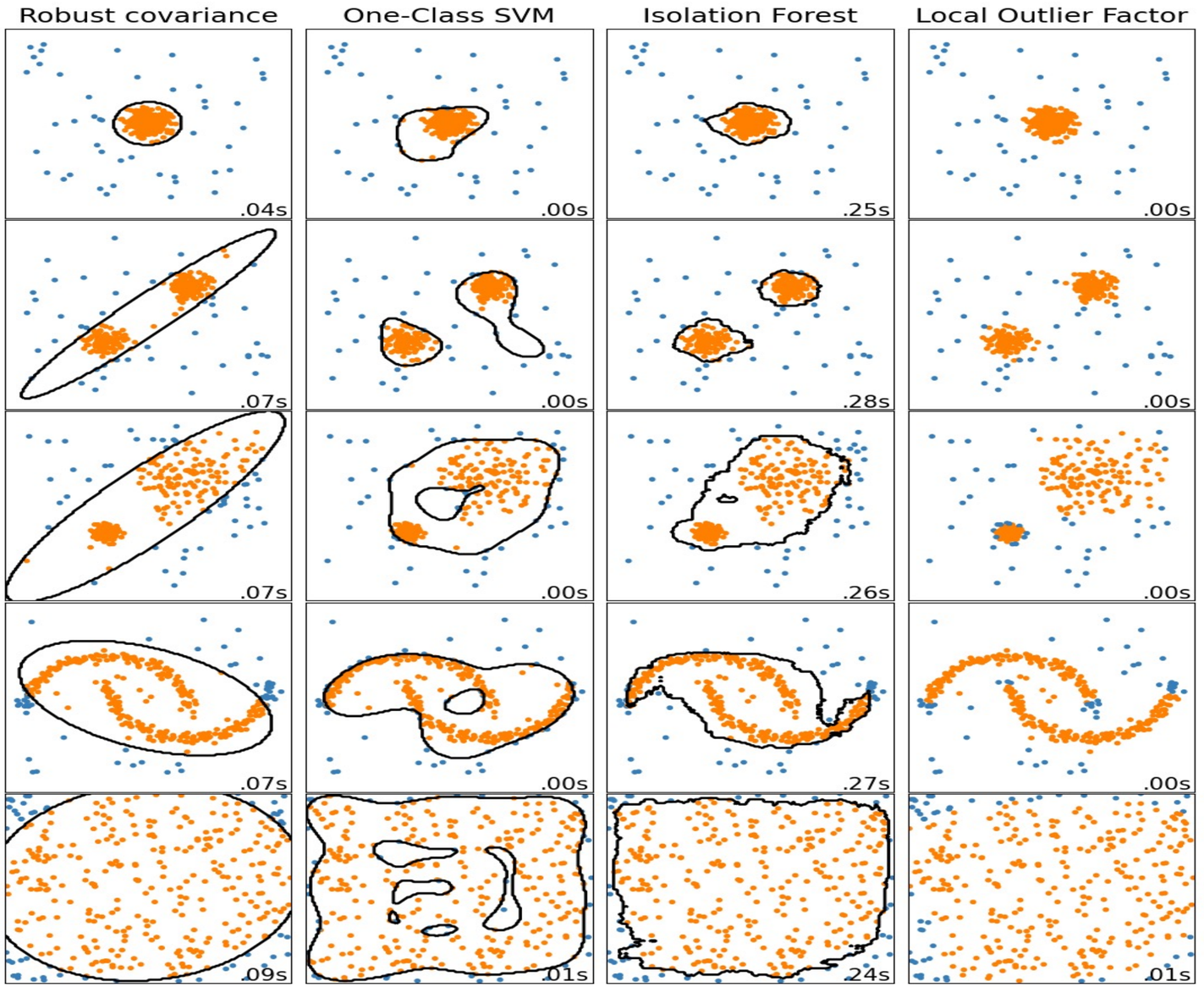
Как их отлавливать?
Isolation Forest



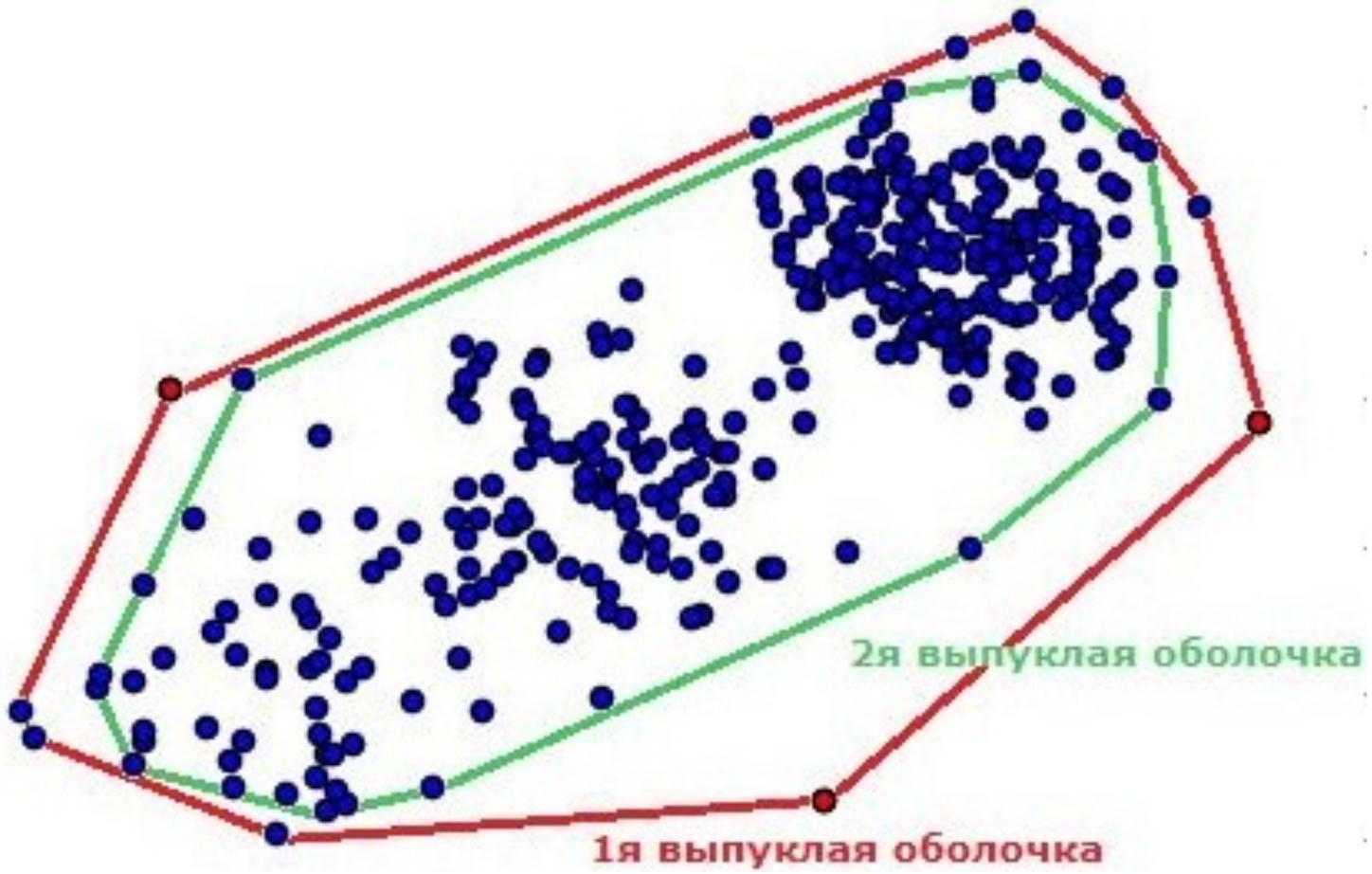
Как их отлавливать?
DBSCAN



Как их отлавливать? Пример нахождения выбросов



Как их отлавливать?



Генерация новых признаков

Создание вещественные признаков

1. деформация (функция над признаком)
2. нормировка (специальный вид деформации)
3. новые признаки (функции над несколькими)
4. дискретизация (binning)

Генерация новых признаков

Кодирование категориальных признаков

1. LabelEncoding
2. Count Encoding
3. OneHotEncoding
4. TargetEncoding
5. CategoryEmbedding

Генерация новых признаков

Создание категориальных признаков

1. конъюнкция признаков
2. создание новых признаков по контекстным
3. экспертное кодирование
4. случайное кодирование

Генерация новых признаков

Временные признаки

1. характеристика момента времени
2. циклические признаки
3. взаимодействие пары признаков

ПРАКТИКА

Спасибо за внимание!