Занятие № 10

Поиск выбросов и генерация новых признаков

Содержание

Что такое выброс?

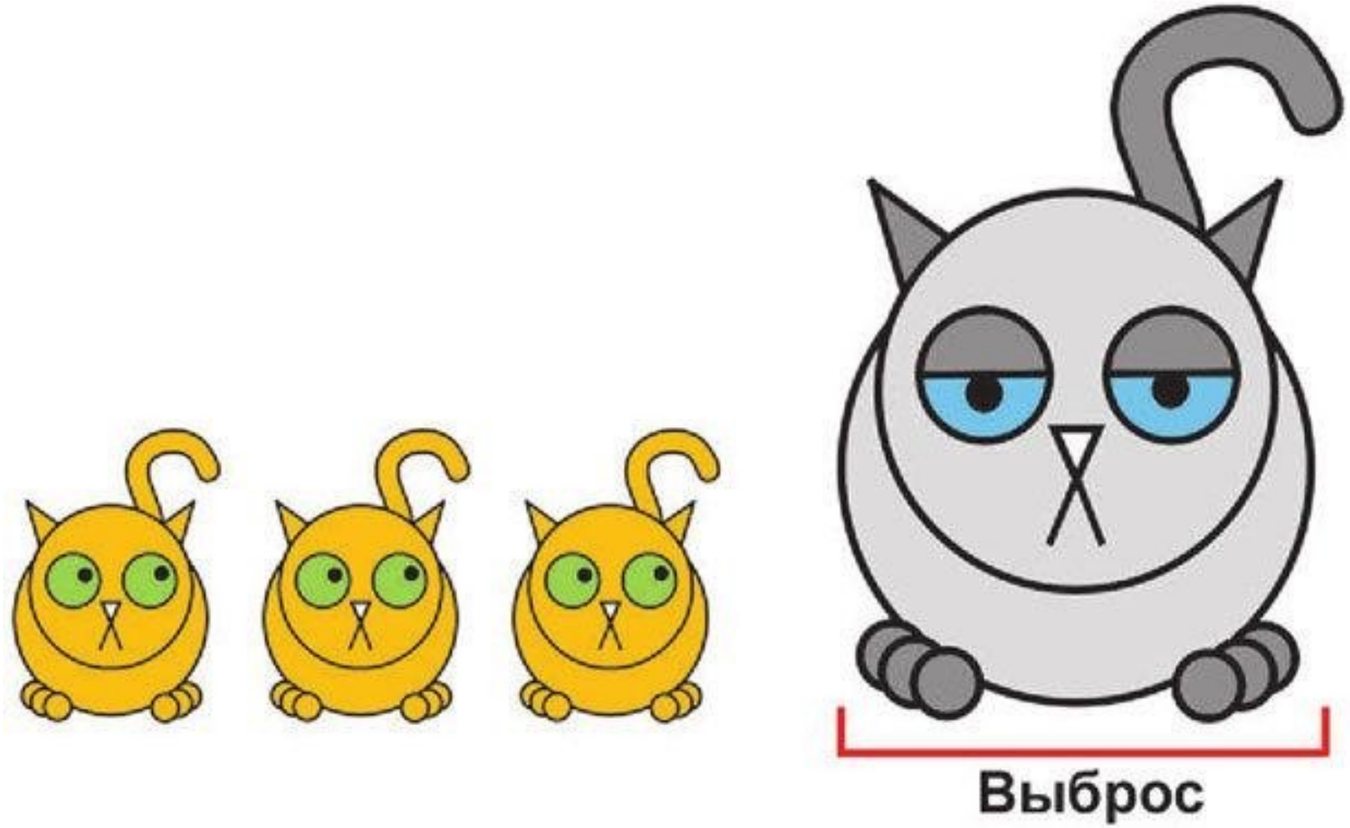
Выбросы - точки данных, которые не пренадлежат определенной популяции

Причины:

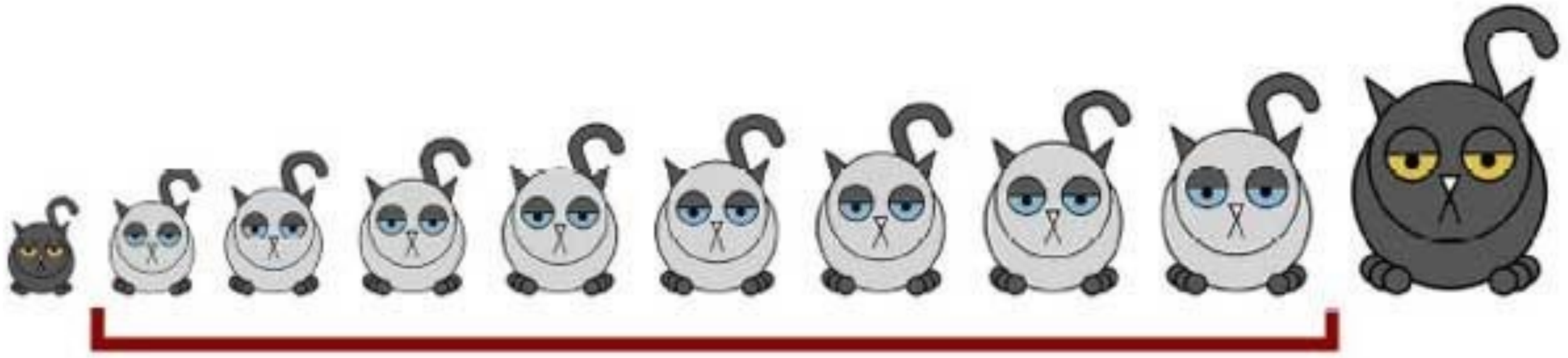
- ошибок в данных
- наличия шумовых объектов
- присутствия объектов «других» выборок

Что такое выброс?

Результат измерения, не подпадающий под общее распределение



Что такое выброс?



Котики для усеченного среднего

Как их отлавливать?

- **Статистические тесты**
  Стандартное отклонение
  Боксплот

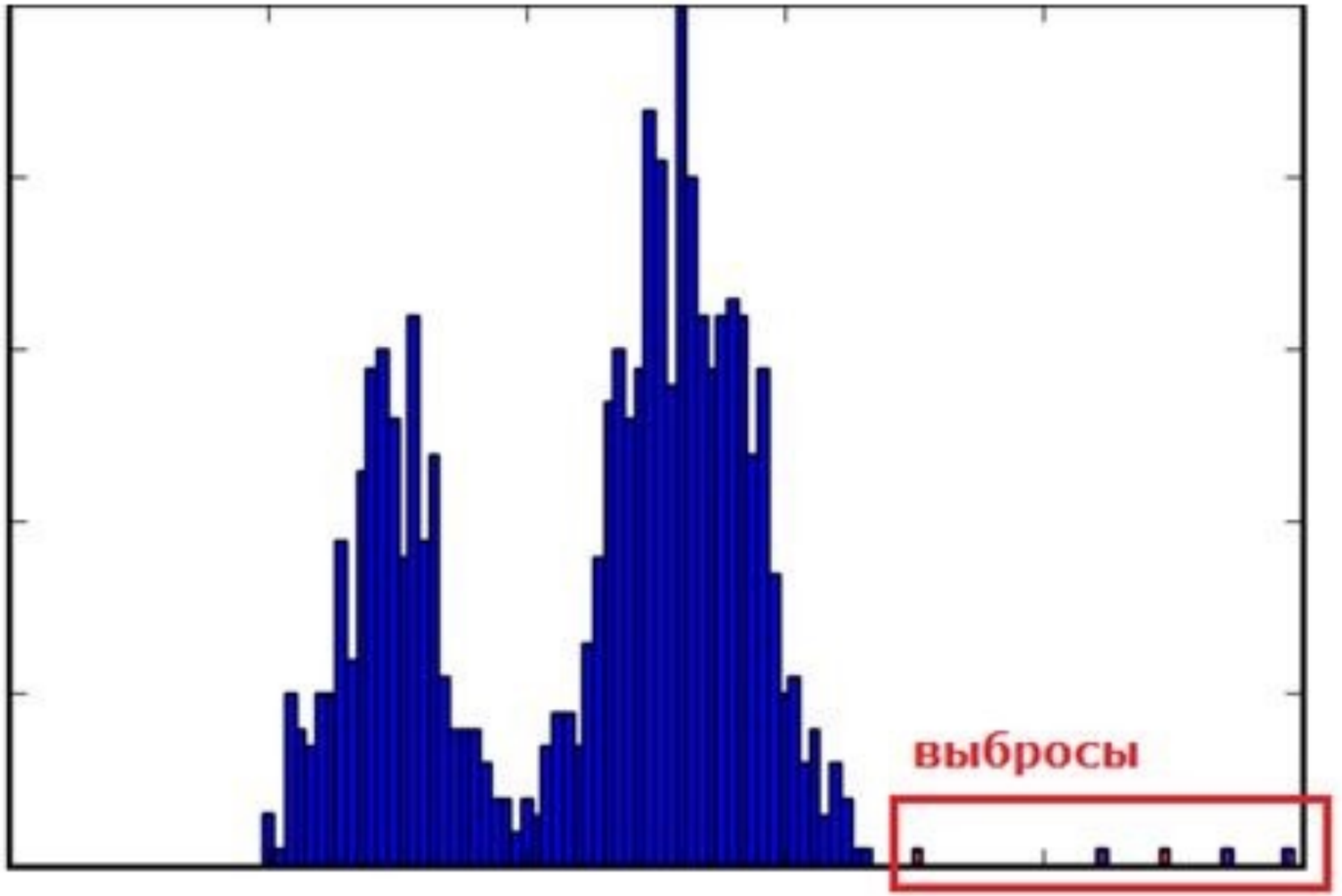
- **Модельные тесты**
  построение модель, которая описывает данные (выбросы плохо описываеются моделью)

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

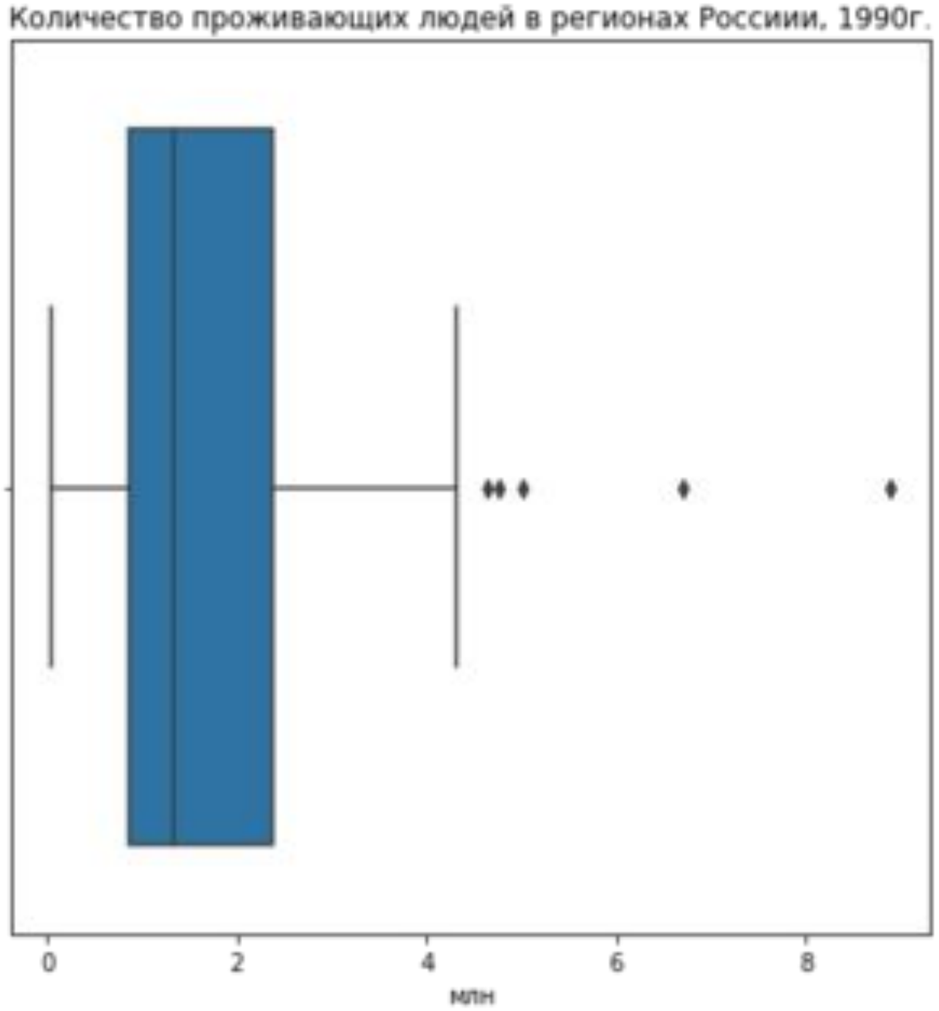
- **Методы машинного обучения**
  Метод опорных векторов для одного класса
  Изолирующий лес

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

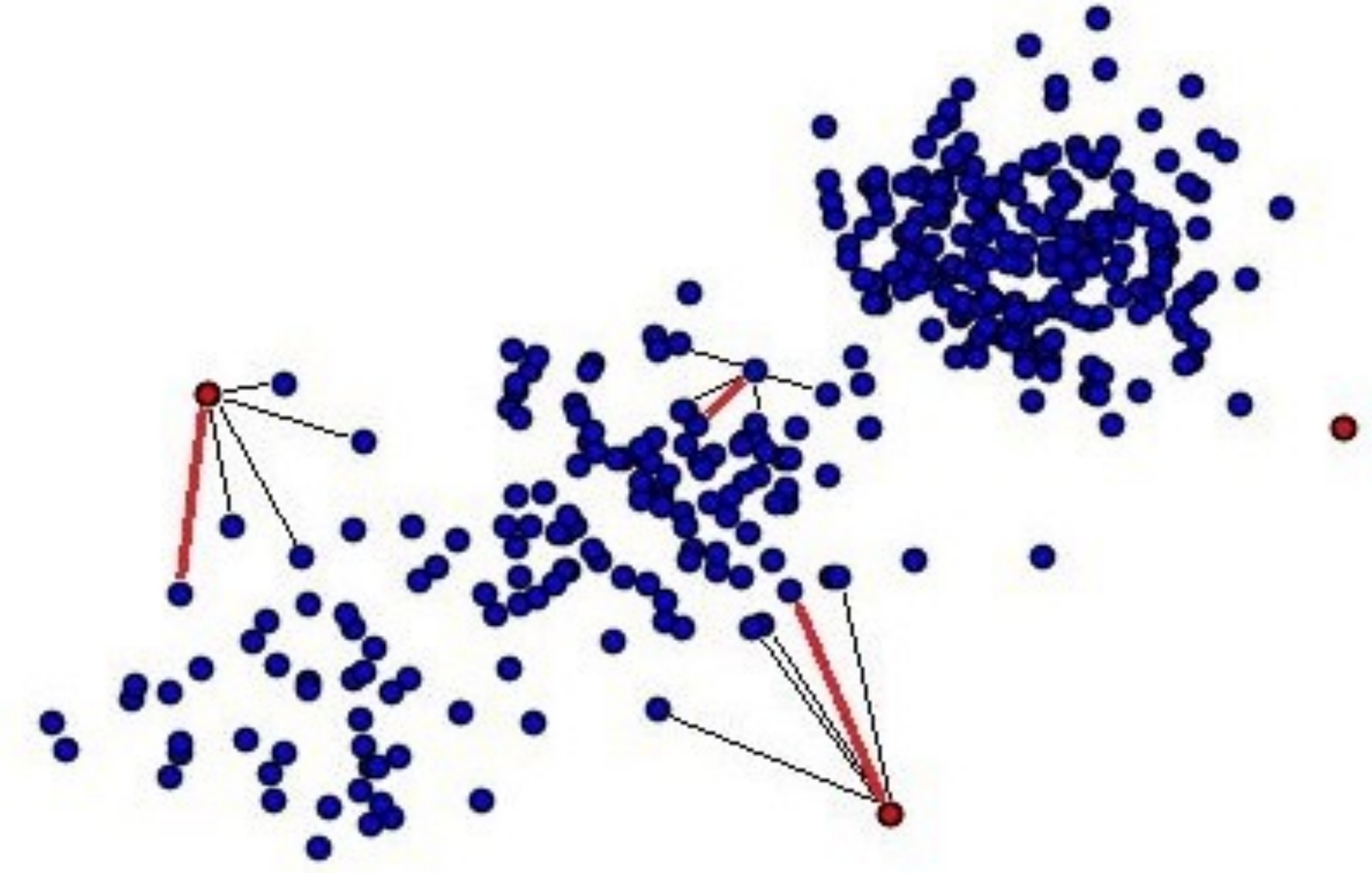
Как их отлавливать?
Выбросы из выборки



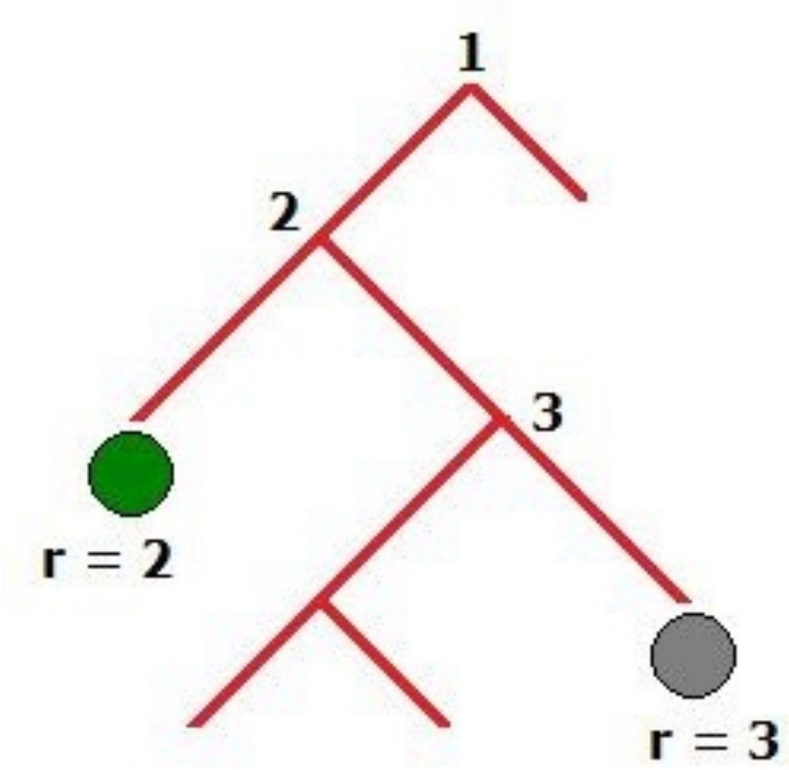
Как их отлавливать?
Выбросы из выборки



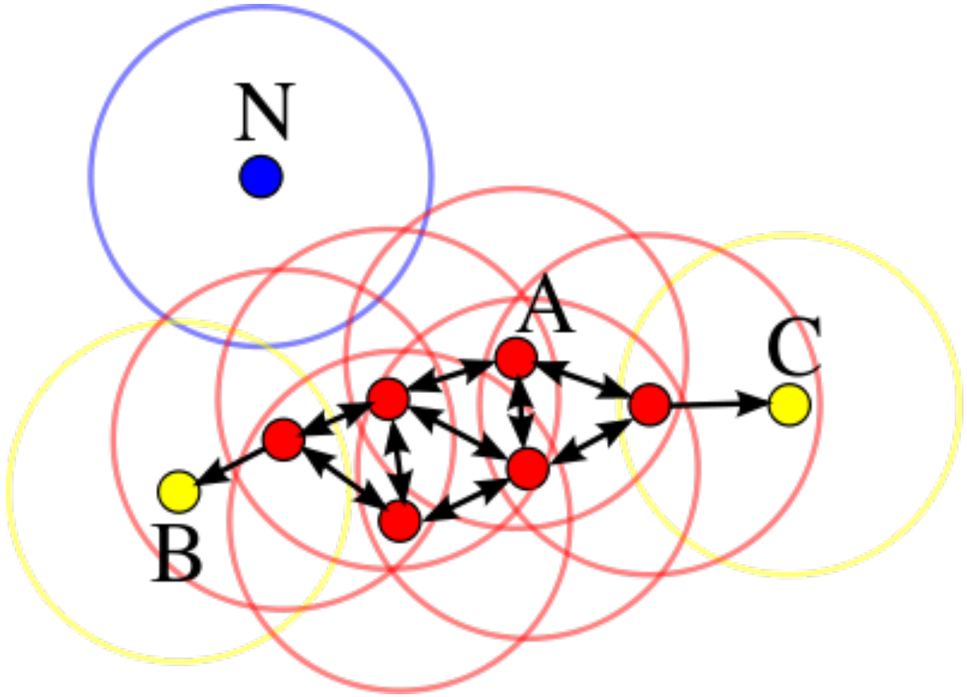
Как их отлавливать?
KNN (LOF)



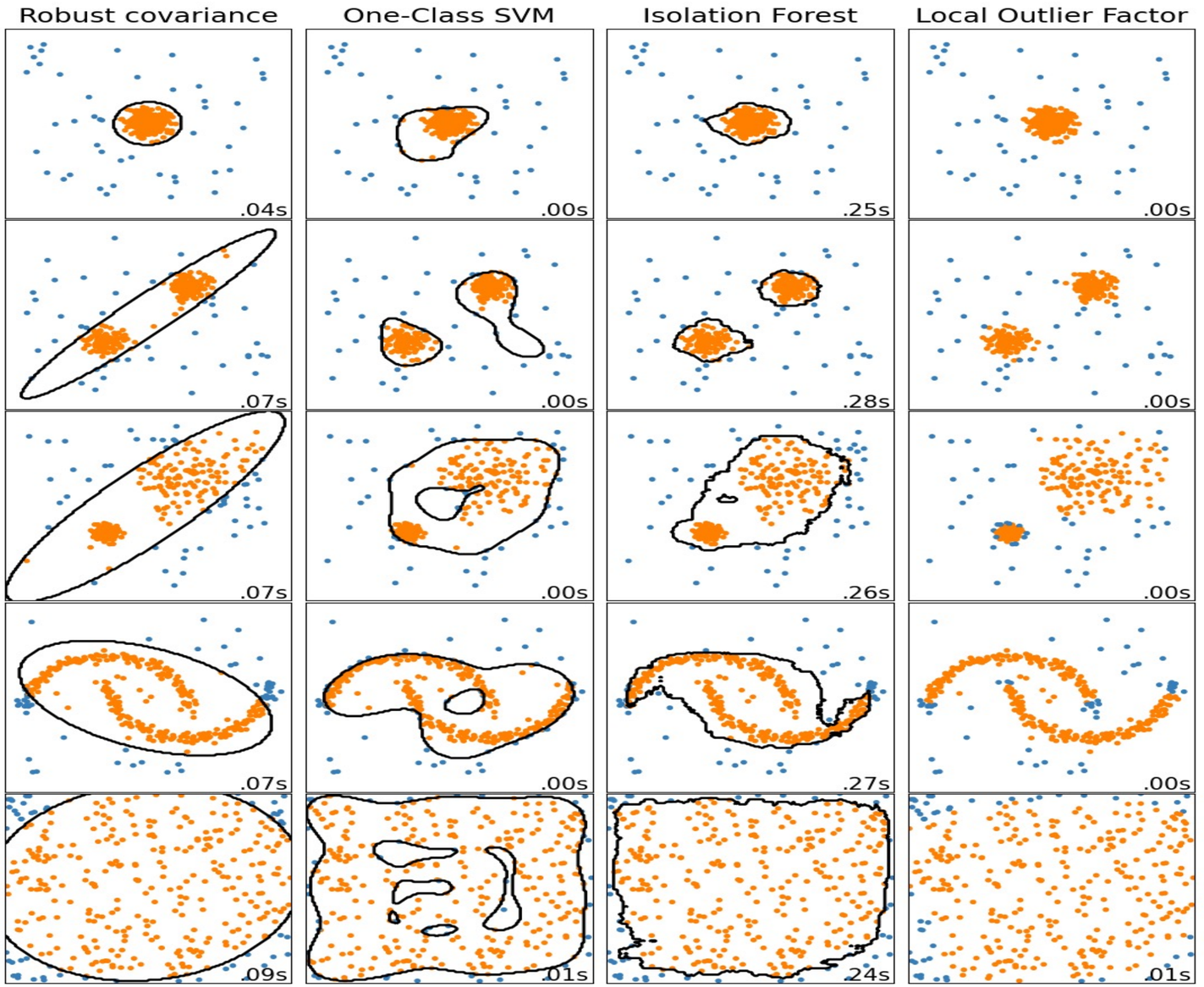
Как их отлавливать?
Isolation Forest



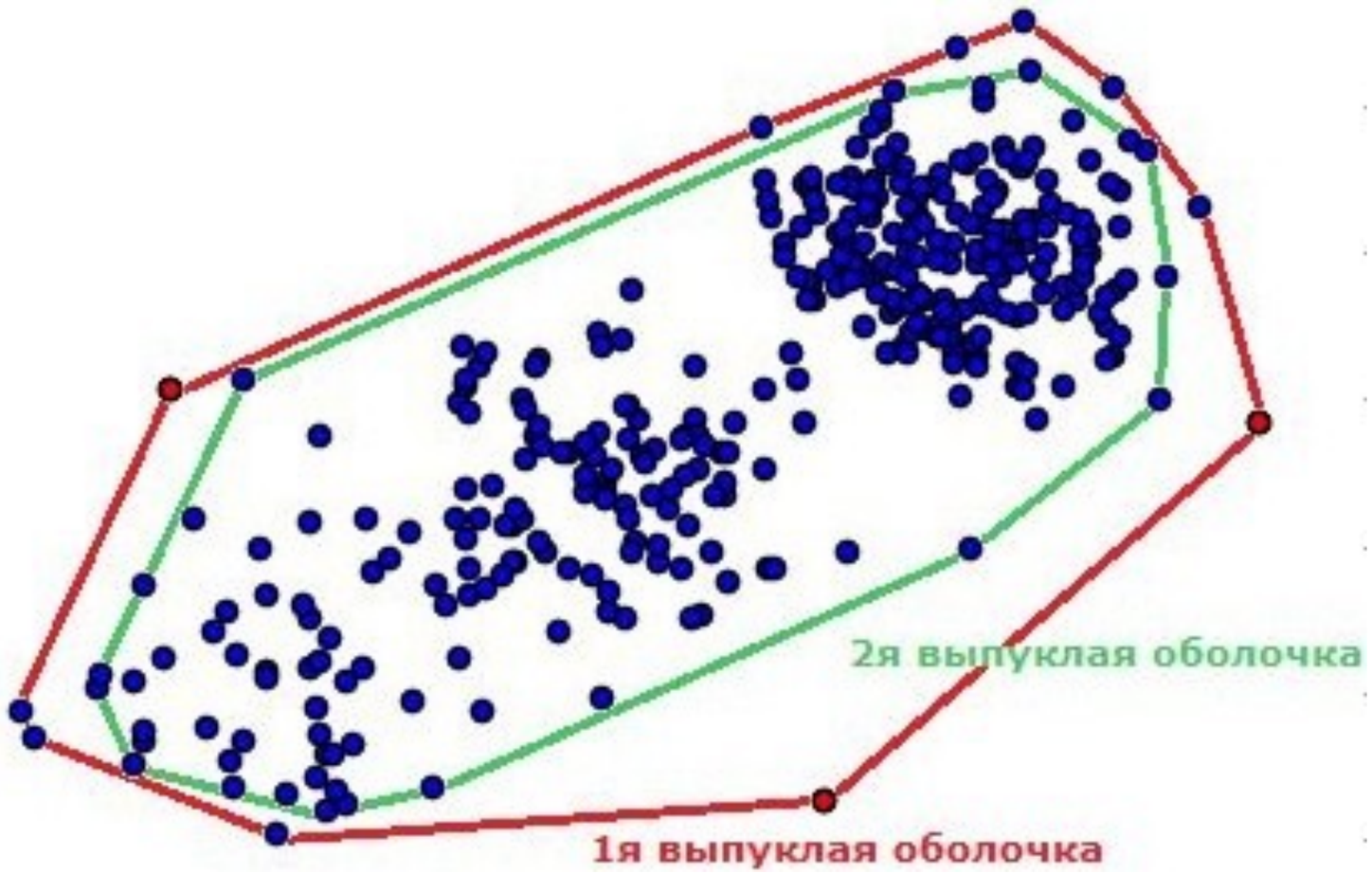
Как их отлавливать?
DBSCAN



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!