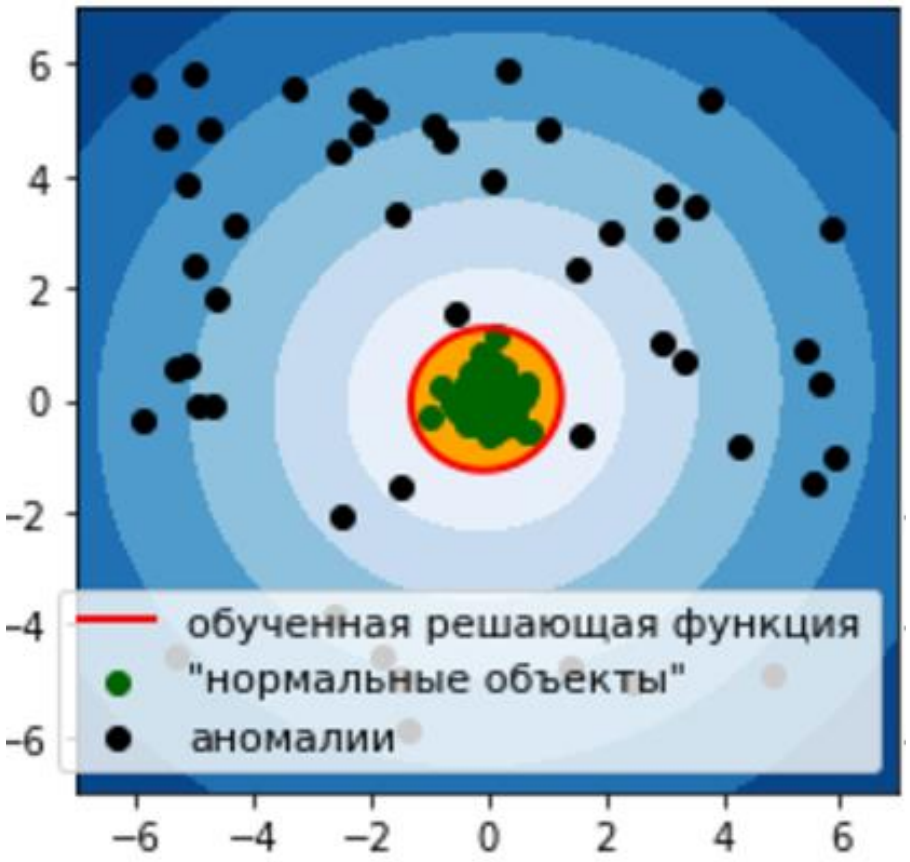
Выявление аномалий

Актуальность задачи

Аномалия (выброс) – пример, событие, которое не соответствуют ожидаемому поведению.



Примеры задач:

- выявить неожиданное поведение рынка, например
  - появление конкурента на рынке,
  - рекламные акции,
  - падение прибыли в определенном регионе,
- выявить мошенничество
  - в банковских операциях
  - в действиях сотрудников компании
- выявить поломку оборудования, используя анализ звука
- предсказать отток абонентов

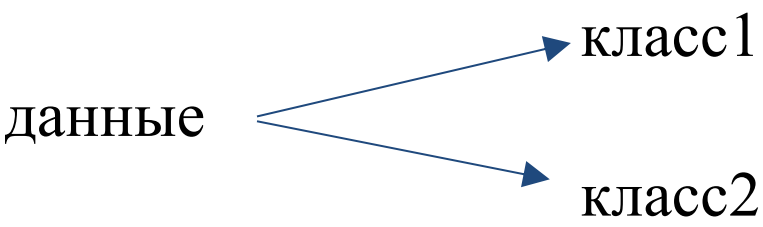
Выявление аномалий как вспомогательная задача:

- найти ошибки ввода информации,
- найти выбросы в данных

Подходы

1. Обучение с учителем

Бинарная классификация



Данные должны быть размечены

1. Легко оценить качество работы детектирования
2. Можно классифицировать аномалии на классы (разные виды аномалий)

2. Обучение без учителя

- статистический метод,
- одноклассовая классификация (one-class SVM),
- обнаружение аномалий на основе плотности (Density-Based Anomaly Detection),
- кластерный анализ,
- предсказательная модель,
- метод на основе PCA.

2. Обучение без учителя

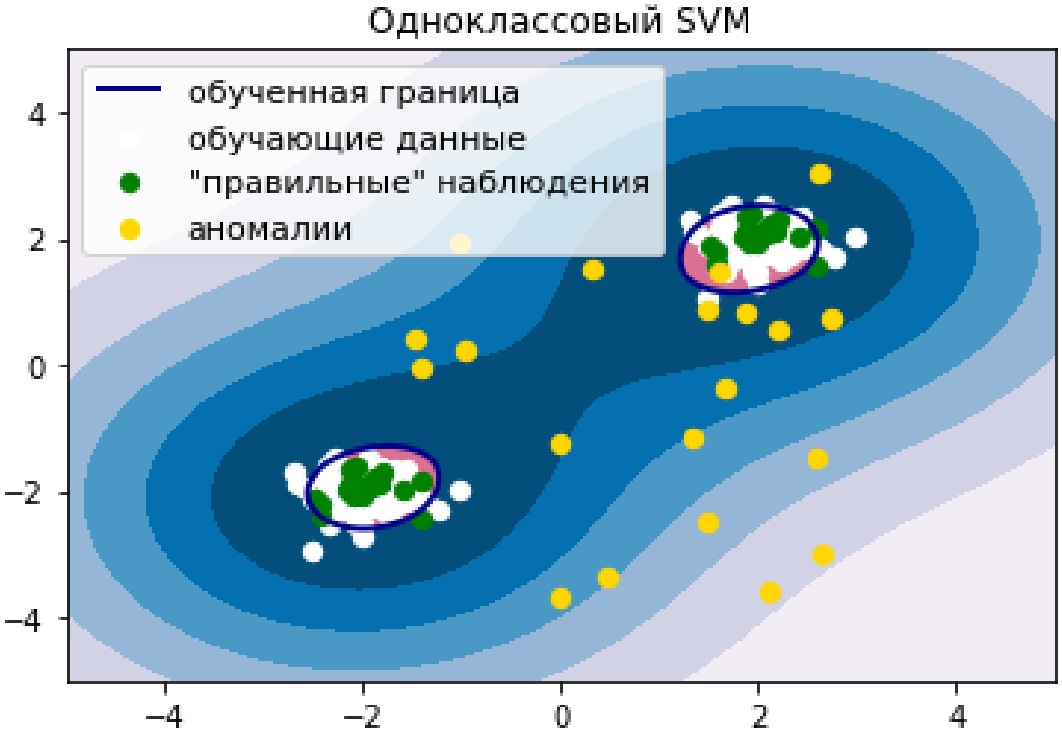
Статистический метод

- анализируем распределение и его параметры
- оцениваем, насколько данный пример соответствует распределению

2. Обучение без учителя

Одноклассовая классификация (One-class SVM)

- определяем нечеткие границы множества
- все, что выходит за границы - аномалии

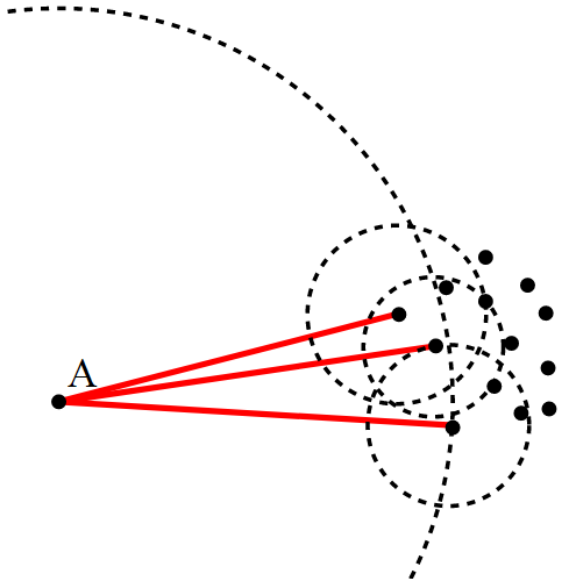


2. Обучение без учителя

Обнаружение аномалий на основе плотности (Density-Based Anomaly Detection)

- Предположение: “нормальные” точки расположены в тесных соседствах, аномалии - далеко от них
- Сравниваем плотности окружения рассматриваемой точки и точек соседей
- Точки с более низкой плотностью - аномалии

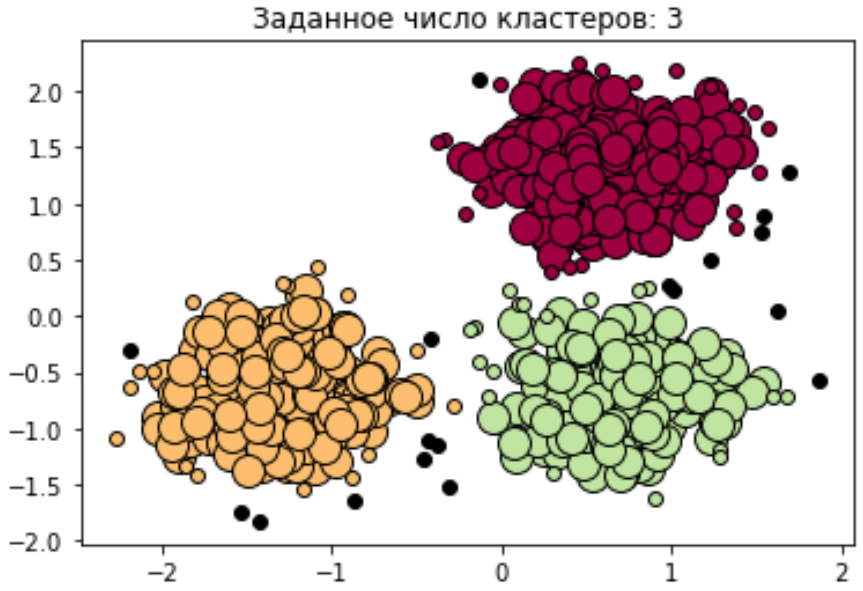
- не требует предварительного обучения
- не нужно хранить в памяти одновременно весь массив данных
- в одном месте это может быть аномалия, в другом - “нормальная” точка



2. Обучение без учителя

Кластерный анализ

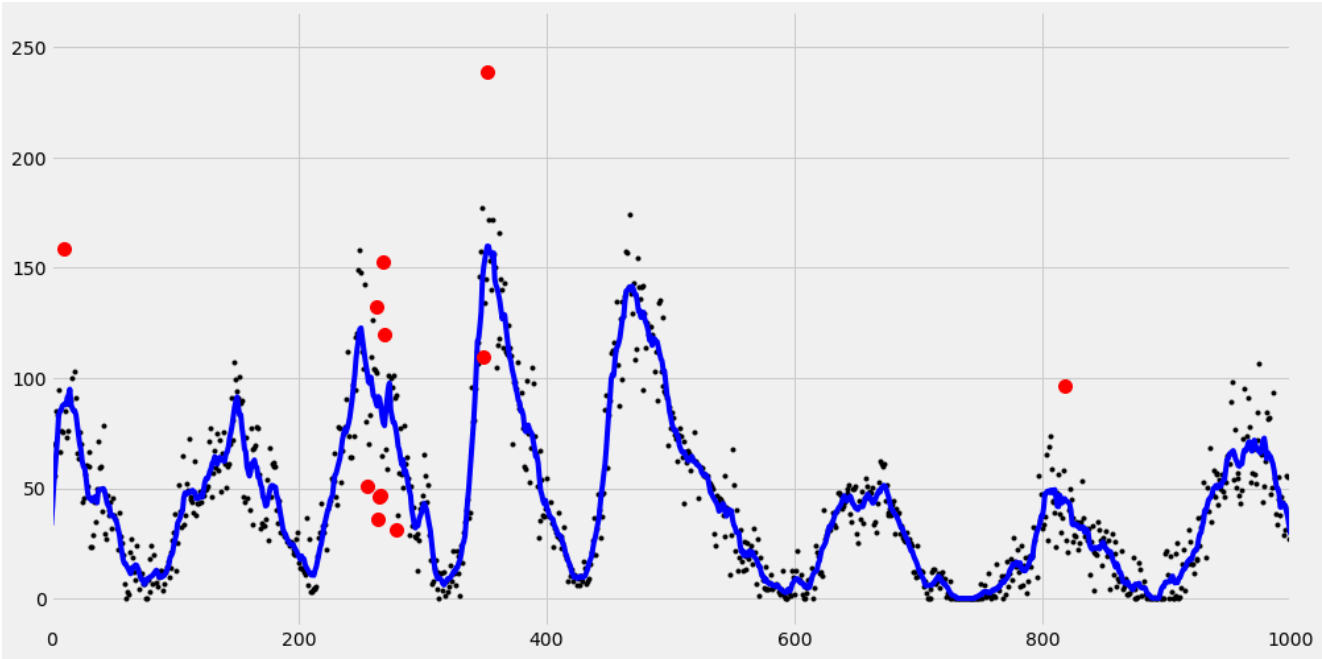
- Основан на предположении: точки, обозначающие похожие элементы, относятся к одному кластеру, что определяется их расстоянием от центра кластера.
- Классический алгоритм - K-means. Создается k кластеров данных. Данные, расположенные далеко от центров кластеров - аномалии (на графике - черные точки).



2. Обучение без учителя

На основе предсказательной модели

- обучаем модель предсказывать (например временной ряд),
- сравниваем результат предсказания с фактом,
- объекты, чье поведение сильно отличается от предсказанного - аномалии.



2. Обучение без учителя

Метод на основе PCA

- выявляет внутреннюю структуру данных
- снижает размерность пространства признаков, выделяя ключевые признаки элементов “нормального” класса
- используем метрику расстояния, чтобы определить аномалии