Лекция 9
Поиск аномалий.

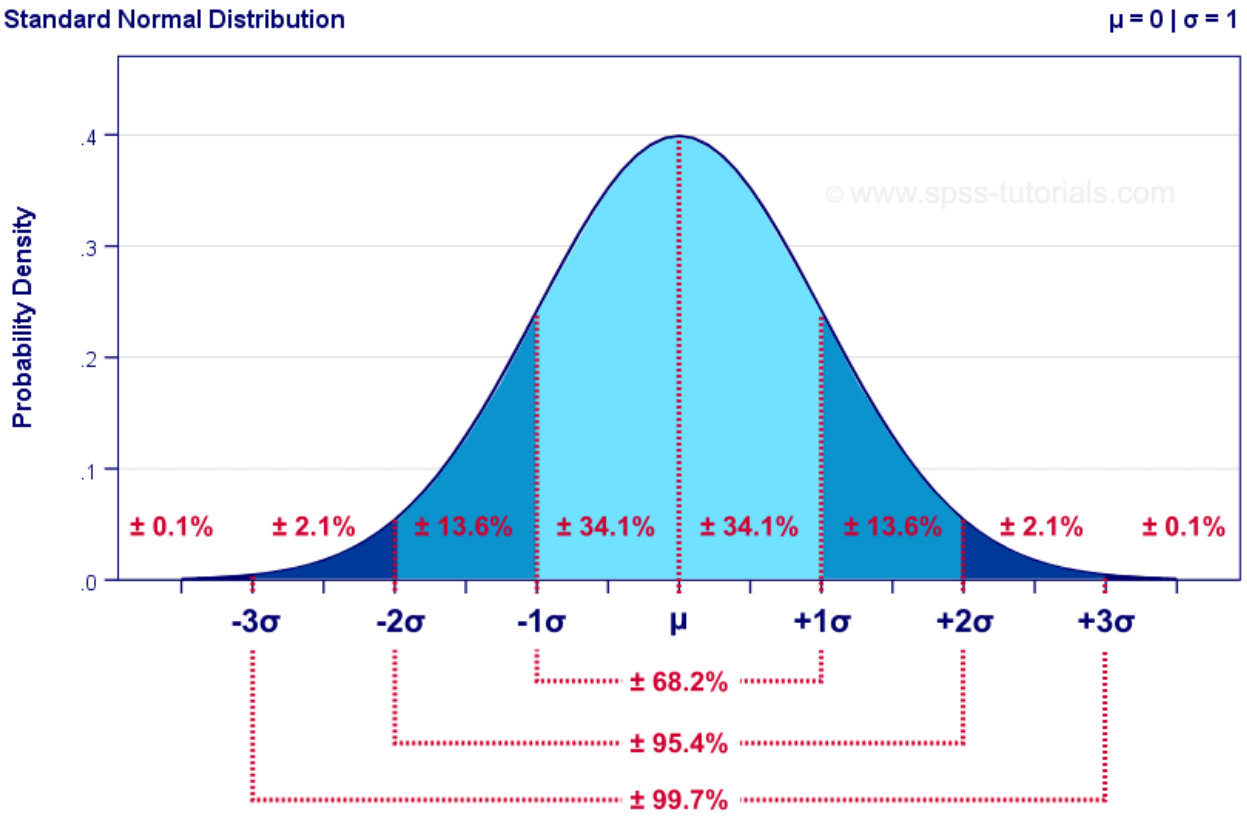
Кантонистова Е.О.

ВШЭ, 2020

Z-SCORE

Если данные распределены нормально, то большинство измерений находится в диапазоне $(m - 3\sigma; m + 3\sigma)$.

Точки, не попавшие в этот диапазон, можно считать выбросами.

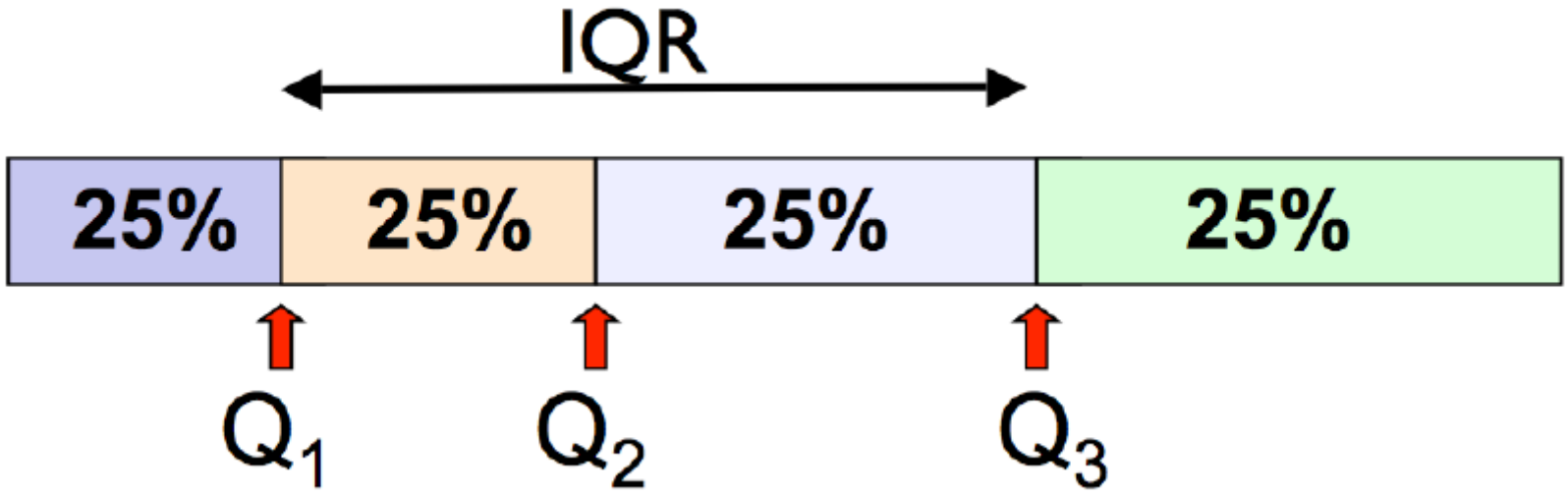


НАХОЖДЕНИЕ ВЫБРОСОВ В ДАННЫХ

Пусть Q1 – первая (25%) квартиль распределения,

Q3 – третья (75%) квартиль распределения.

- Величина IQR = Q3 – Q1 называется ***интерквартильным размахом***.



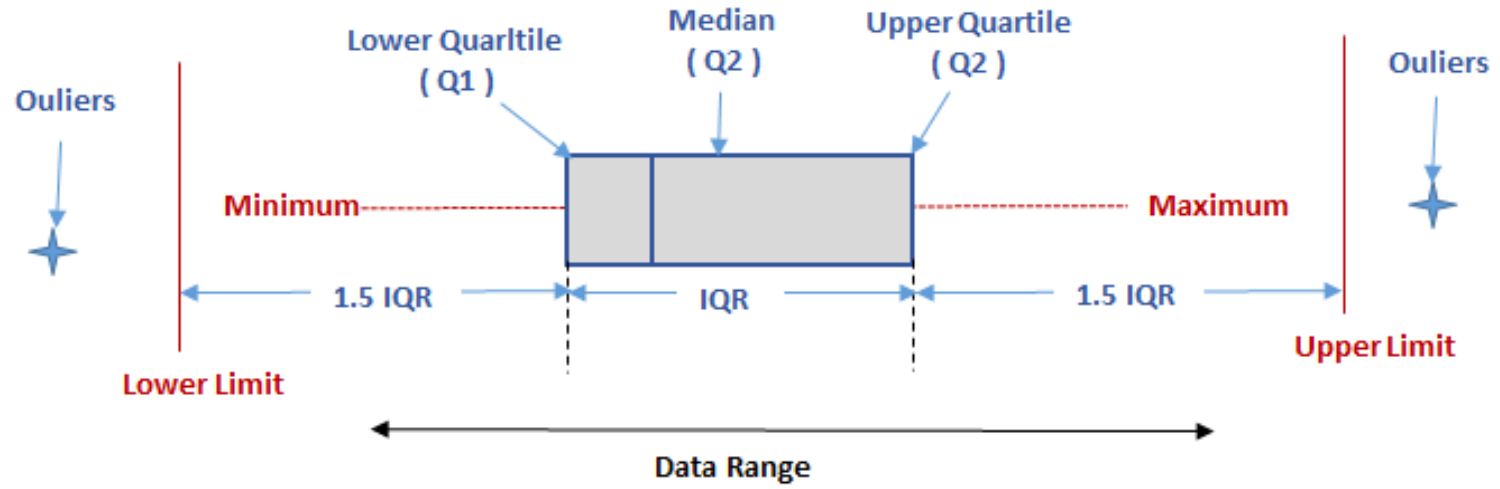
НАХОЖДЕНИЕ ВЫБРОСОВ В ДАННЫХ

- **Слабые выбросы** – это значения, которые меньше 25%-квартили минус 1,5*IQR или больше 75%-квартили плюс 1,5*IQR:

$$x < Q1 - 1{,}5 \cdot IQR \text{ или } x > Q3 + 1{,}5 \cdot IQR$$

- **Сильные выбросы** – это значения, которые меньше 25%-квартили минус 3*IQR или больше 75%-квартили плюс 3*IQR:

$$x < Q1 - 3 \cdot IQR \text{ или } x > Q3 + 3 \cdot IQR$$



ПОИСК АНОМАЛИЙ С ПОМОЩЬЮ МОДЕЛЕЙ ML

Идея: можно настроить модель машинного обучения так, чтобы на нормальных объектах она принимала значения, близкие к нулю (или, например, положительные значения). Тогда если прогноз на объекте сильно отличается от прогноза на обучающей выборке, то такой объект можно считать аномальным.

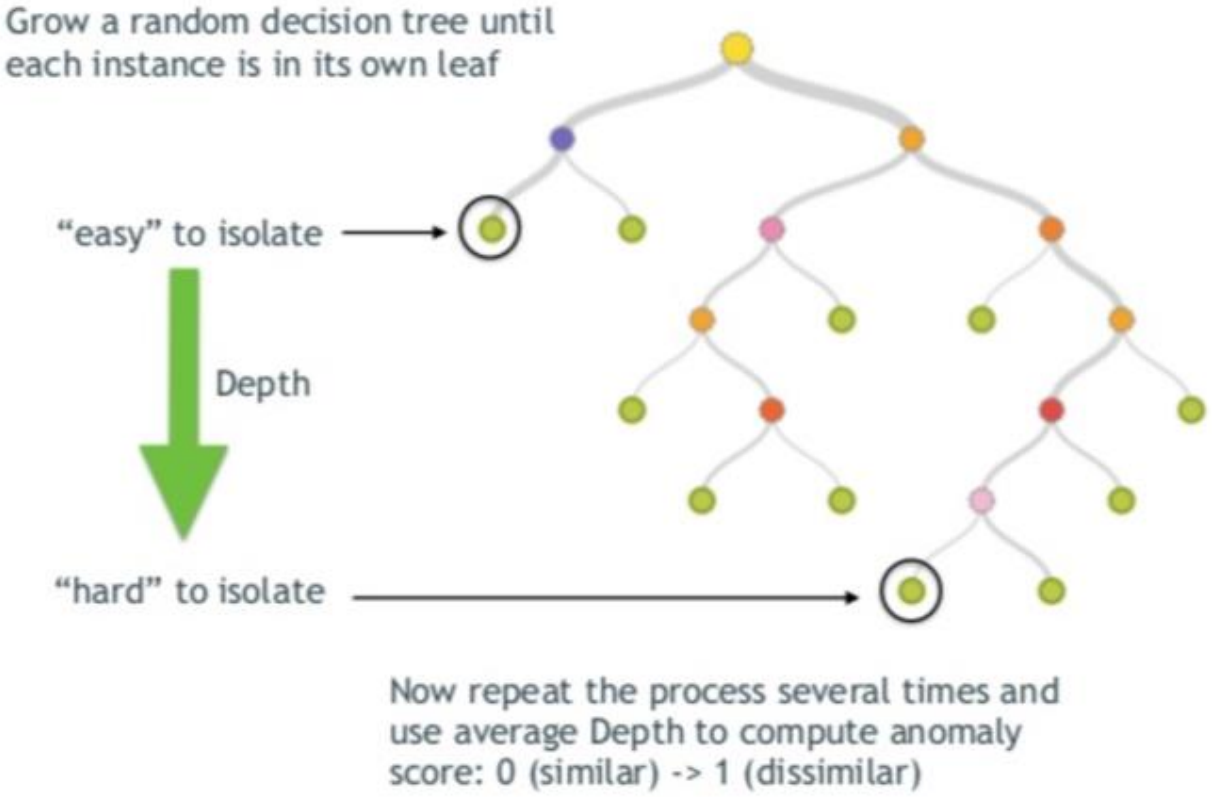
ISOLATION FOREST

- Строим лес, состоящий из N деревьев. Каждый признак и порог выбираем случайно. Останавливаемся, когда в вершине 1 объект или когда построили дерево максимальной глубины.

Идея: чем сильнее объект отличается от большинства, тем раньше он будет отделен от основной выборки случайными разбиениями => выбросы – объекты, которые оказались на небольшой глубине.

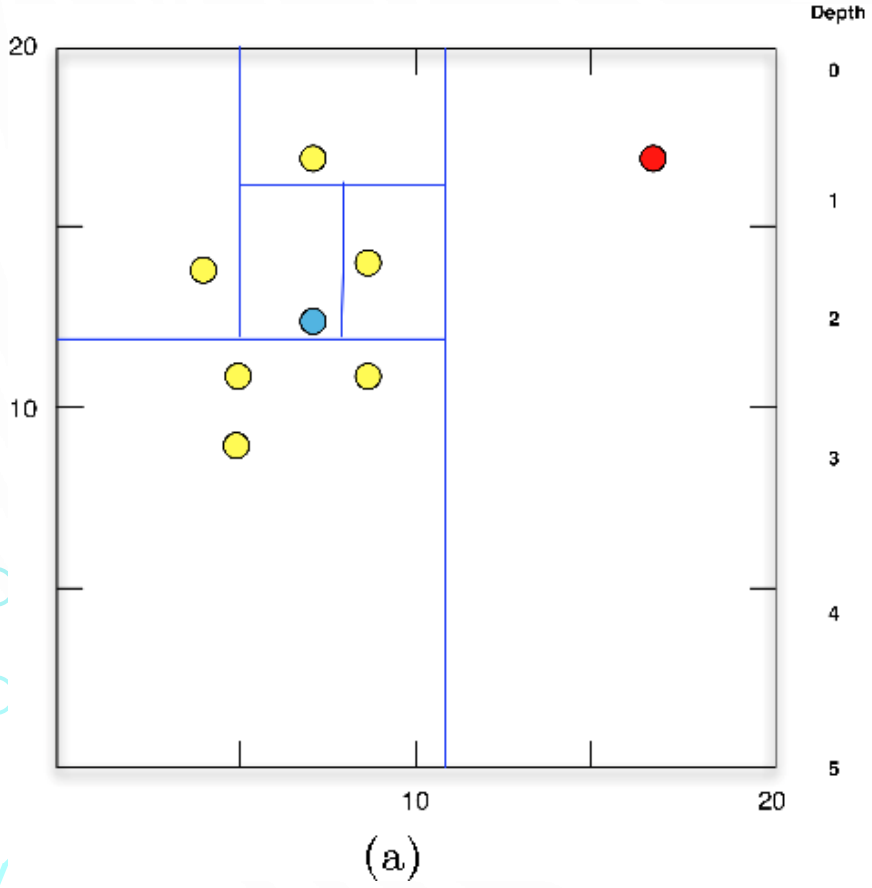
ISOLATION FOREST

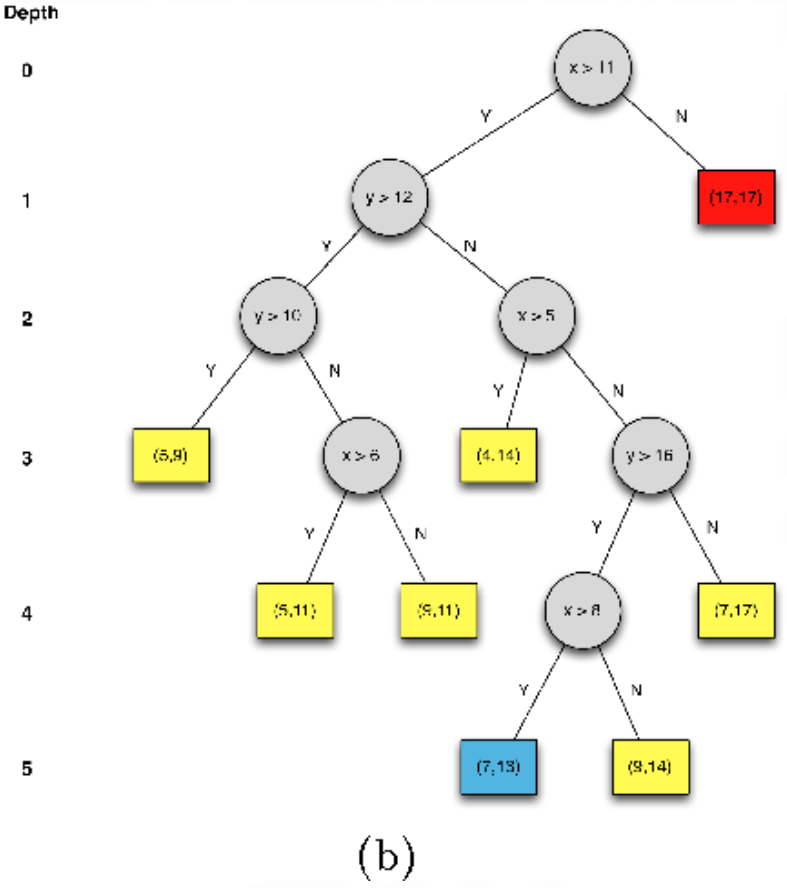
Идея: чем сильнее объект отличается от большинства, тем раньше он будет отделен от основной выборки случайными разбиениями => выбросы – объекты, которые оказались на небольшой глубине.



ISOLATION FOREST

Идея: чем сильнее объект отличается от большинства, тем раньше он будет отделен от основной выборки случайными разбиениями => выбросы – объекты, которые оказались на небольшой глубине.



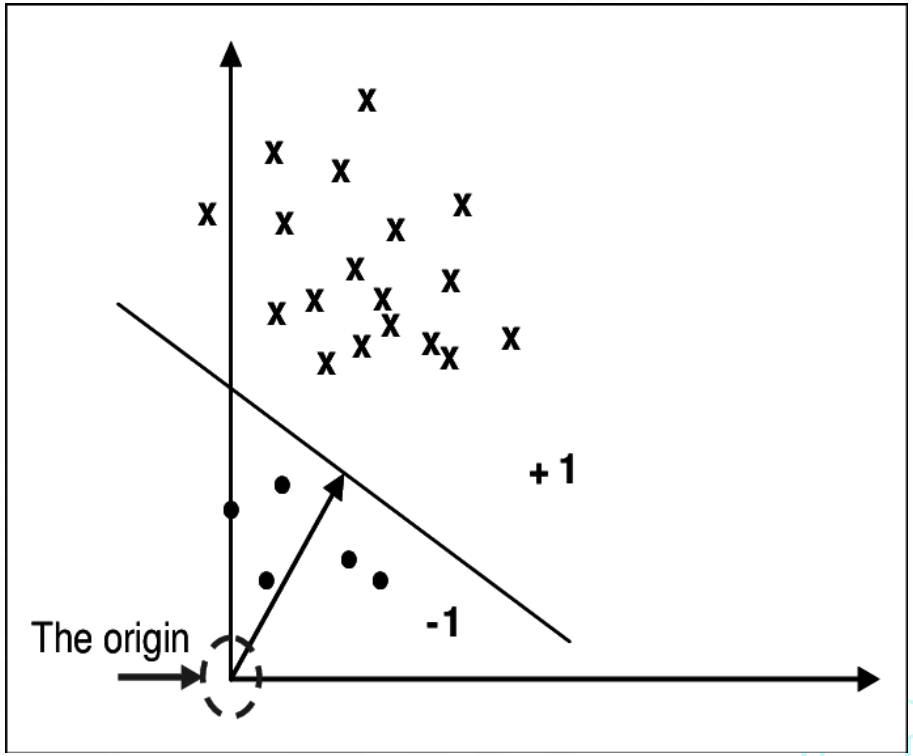


ONE-CLASS SVM

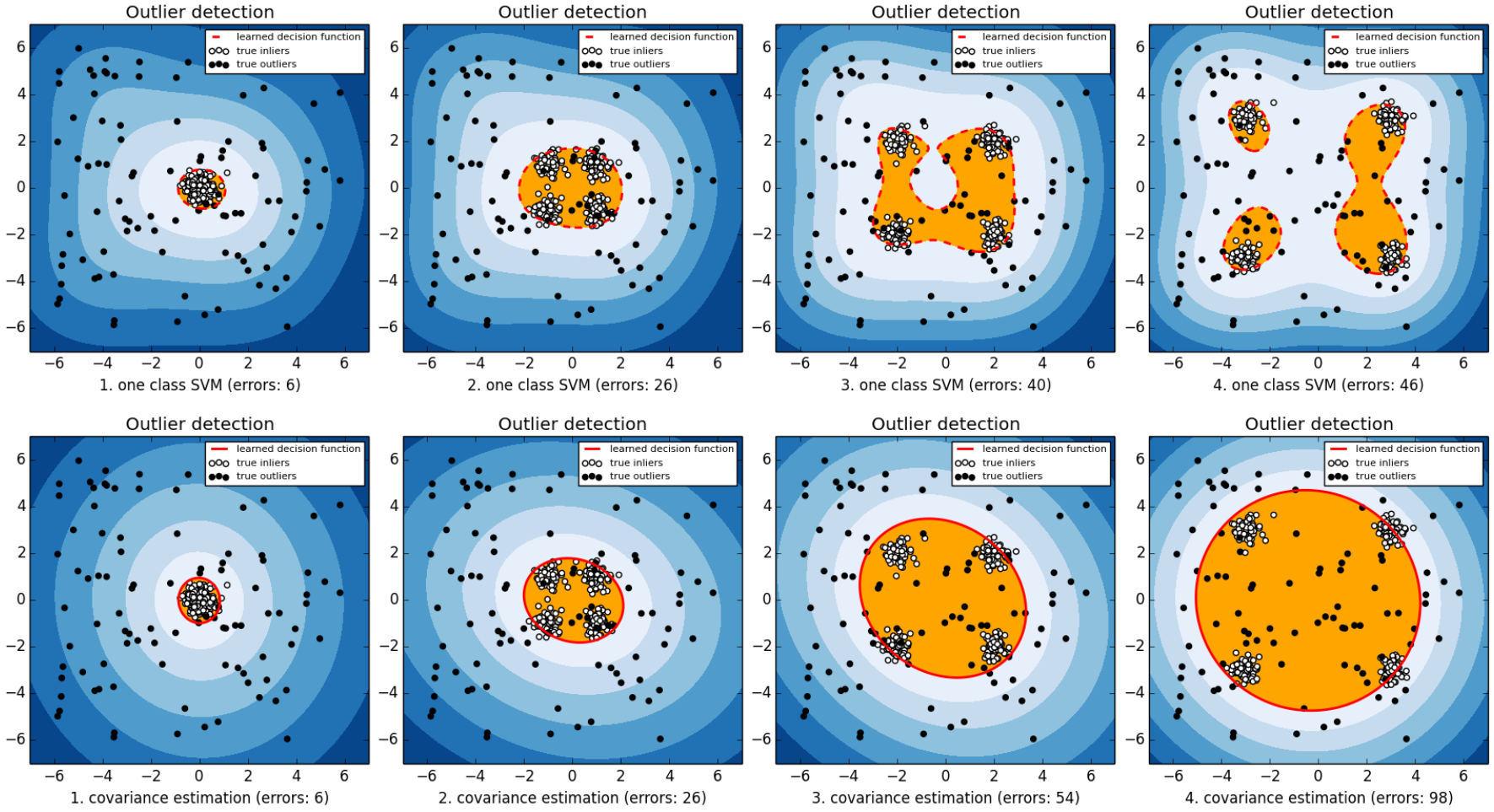
Метод строит линейную функцию $a(x) = sign(w, x)$ так, чтобы она отделяла выборку от начала координат с максимальным отступом, а именно:

- $a(x)$ отделяет как можно больше объектов выборки от нуля
- имеет большой отступ

Тогда объекты с $a(x) = -1$ – это аномалии.

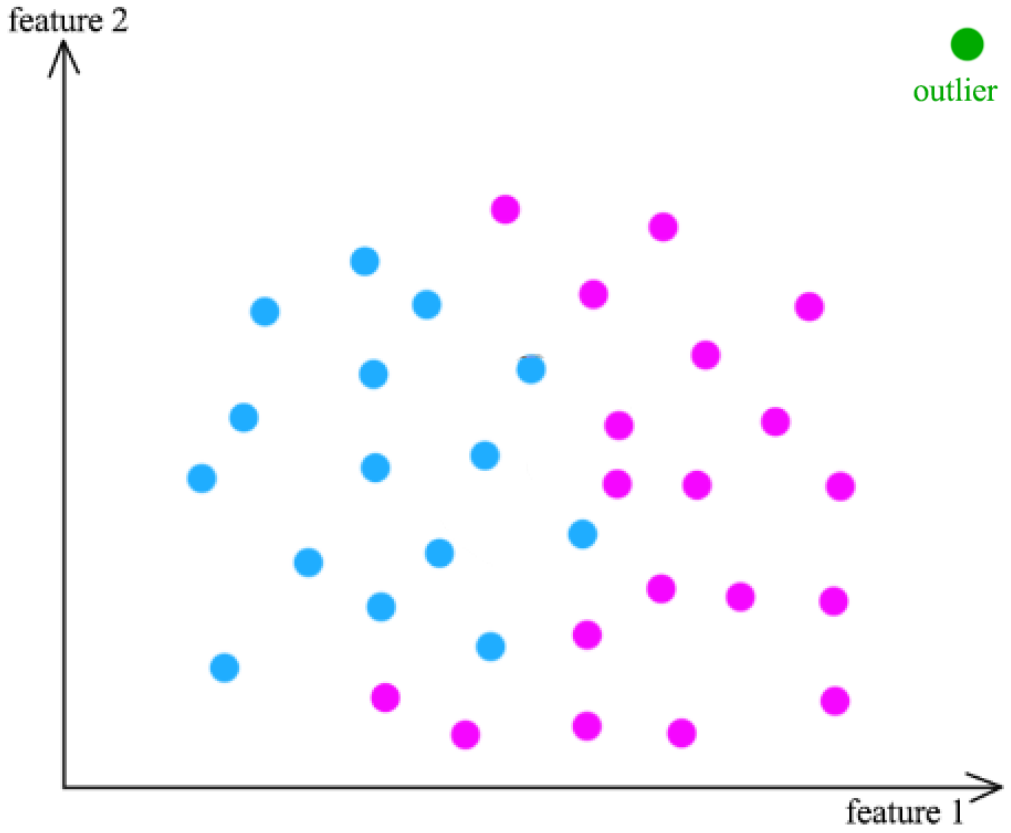


ONE-CLASS SVM С RBF-ЯДРОМ



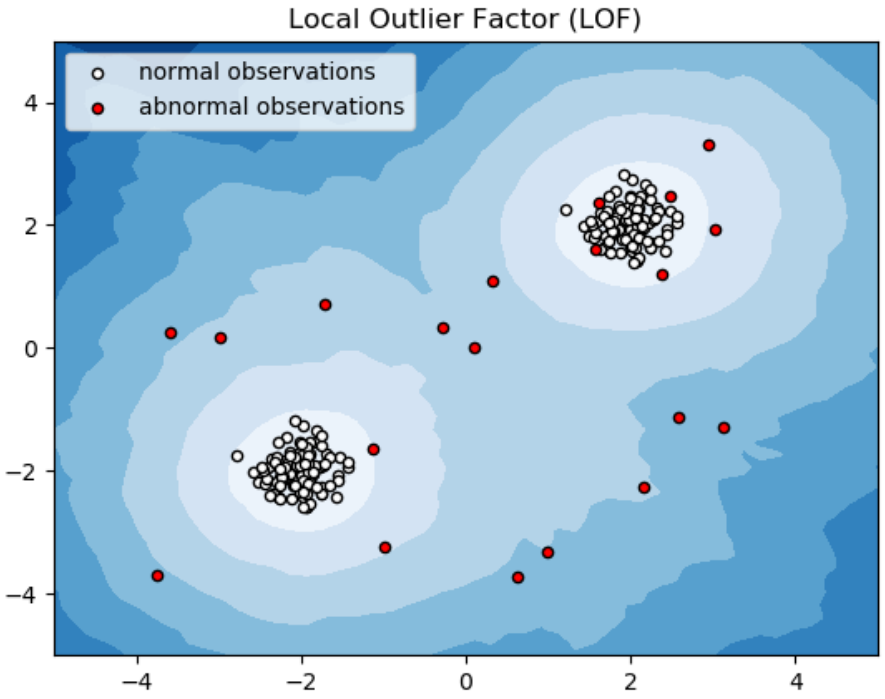
ПОИСК ВЫБРОСОВ С ПОМОЩЬЮ KNN

- Вычисляем среднее расстояние от каждой точки до её ближайших k соседей
- Точки с наибольшим средним расстоянием – выбросы



LOCAL OUTLIER FACTOR

- Задаем плотность распределения в точке, используя k ближайших соседей
- Точки, плотность распределения в которых значительно меньше, чем у соседей – выбросы.



- https://scikit-learn.org/stable/modules/outlier_detection.html
- https://github.com/yzhao062/pyod