Занятие 12
Поиск аномалий.

Кантонистова Е.О.

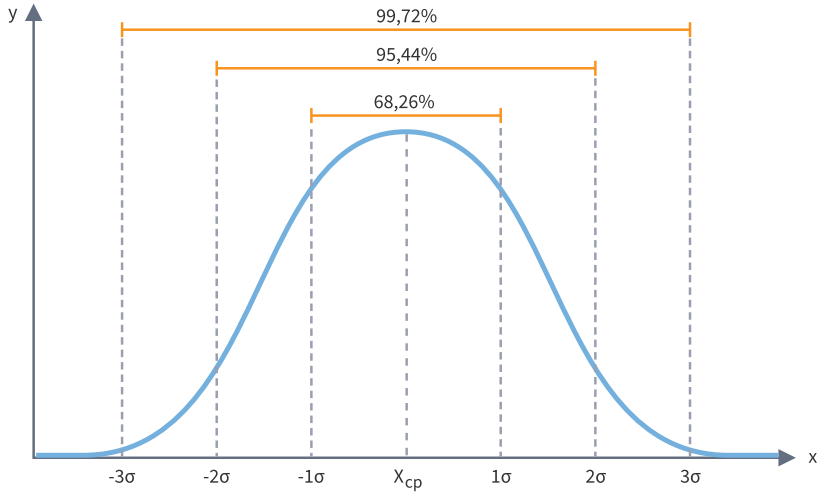
ВШЭ, 2021

РАБОТА С ВЫБРОСАМИ

1. Статистические методы (правило трех сигм, интерквартильный размах).
2. Методы машинного обучения.

1. ПРАВИЛО ТРЕХ СИГМ

- Для случайных величин, распределенных по нормальному закону, вероятность того, что случайная величина отклонится от своего математического ожидания более чем на три стандартных отклонения, практически равна нулю.



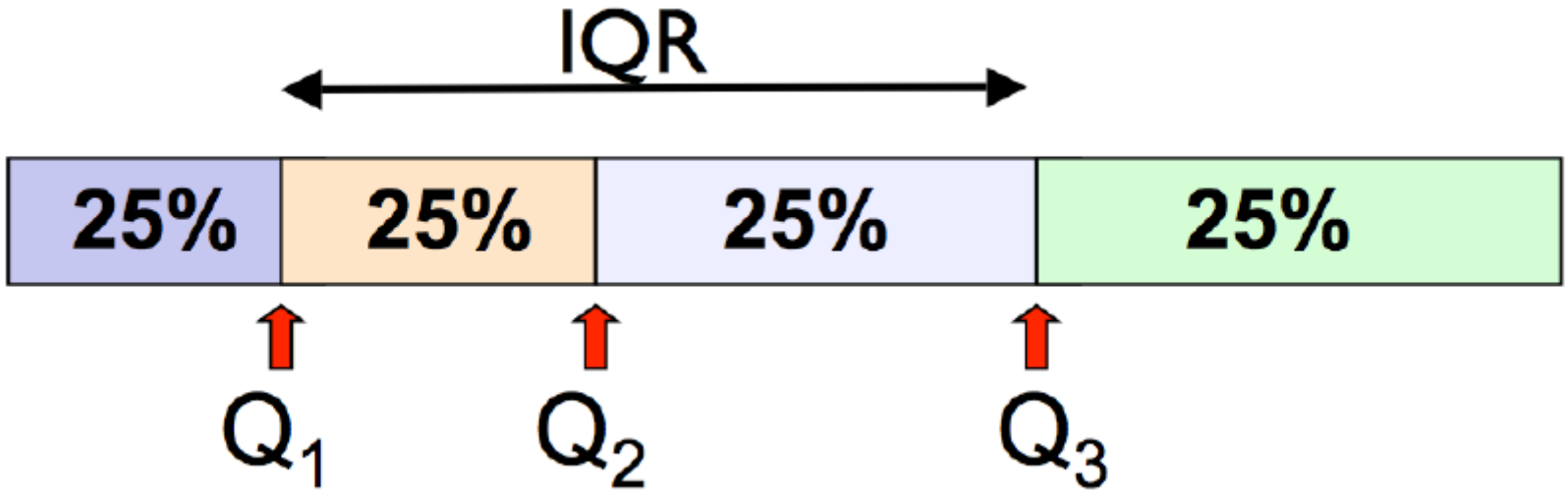
- Выбросами объявляются объекты, имеющие стандартное отклонение $\geq 3\sigma$ от математического ожидания.

2. ИНТЕРКВАРТИЛЬНЫЙ РАЗМАХ

Пусть Q1 – первая (25%) квартиль распределения,

Q3 – третья (75%) квартиль распределения.

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



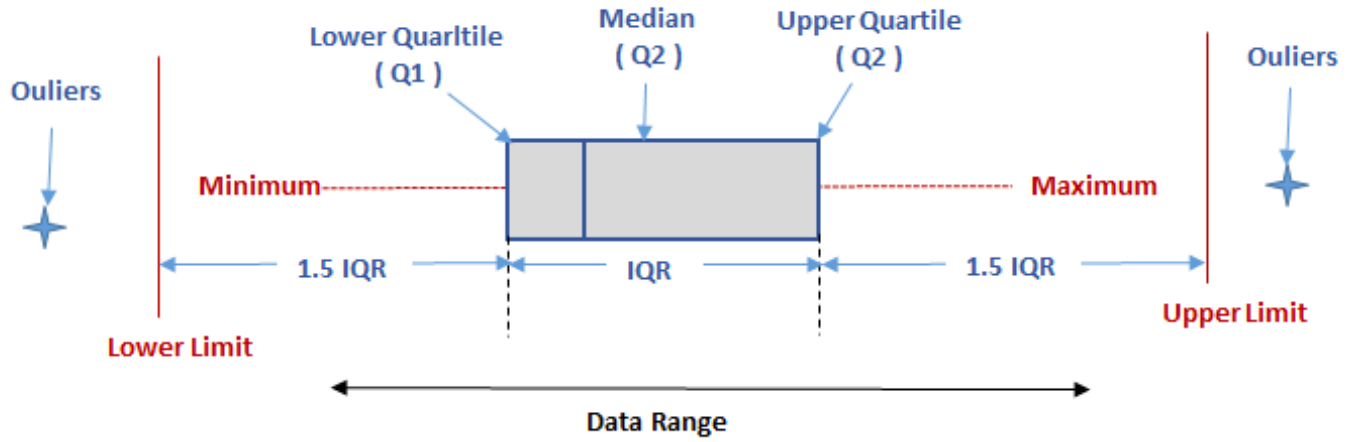
2. ИНТЕРКВАРТИЛЬНЫЙ РАЗМАХ

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

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

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

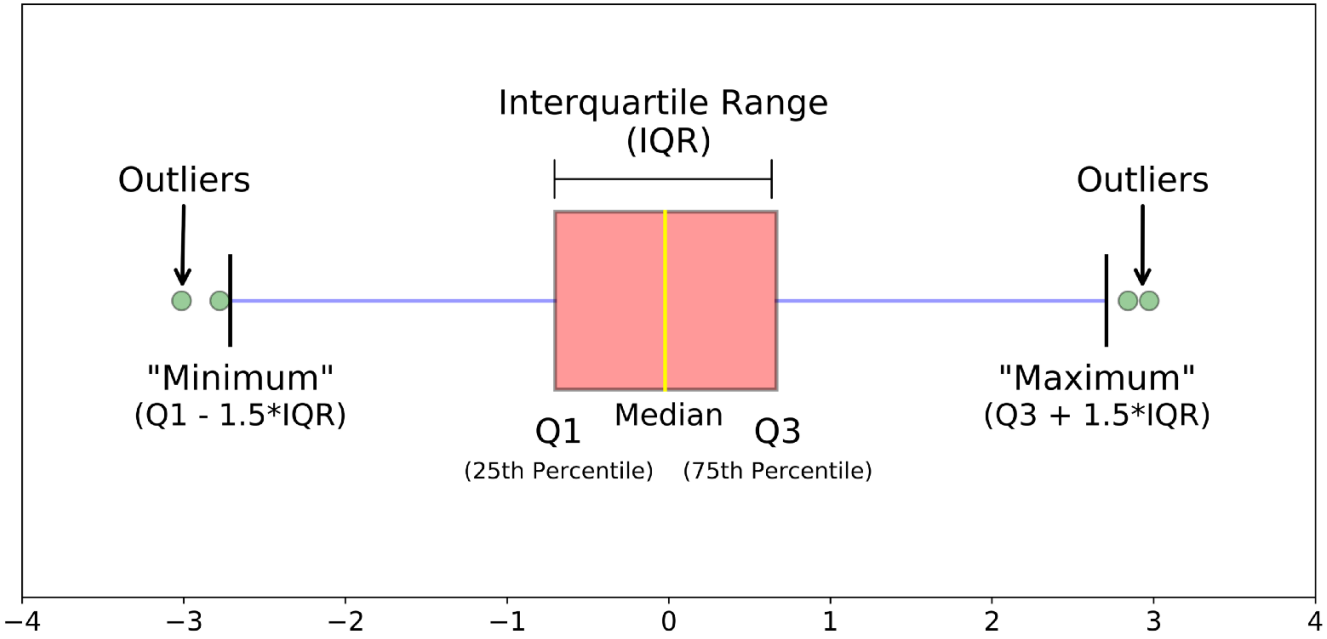
$$x < Q1 - 3 \cdot IQR \text{ или } x > Q3 + 3 \cdot IQR$$



ЯЩИК С УСАМИ

Ящик с усами – это диаграмма, которая показывает:

- одномерное распределение вероятностей (квартили)
- границы попадания “нормальных” точек
- выбросы



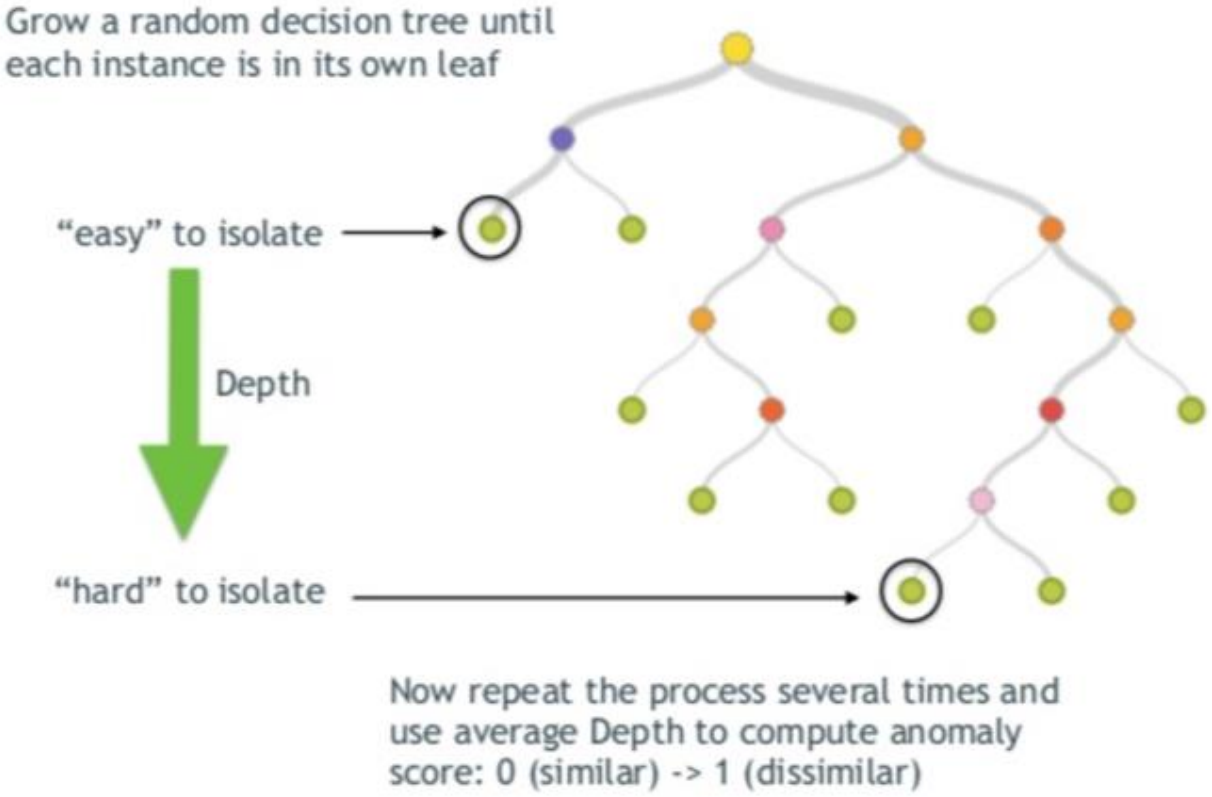
ISOLATION FOREST

- 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.

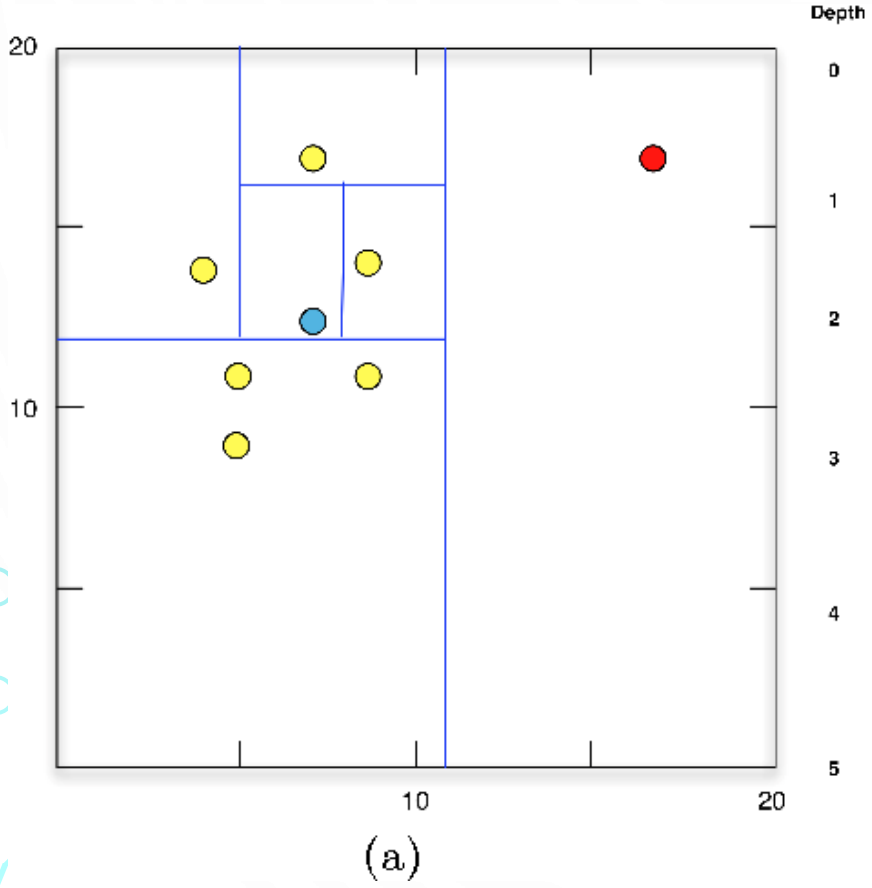
ISOLATION FOREST

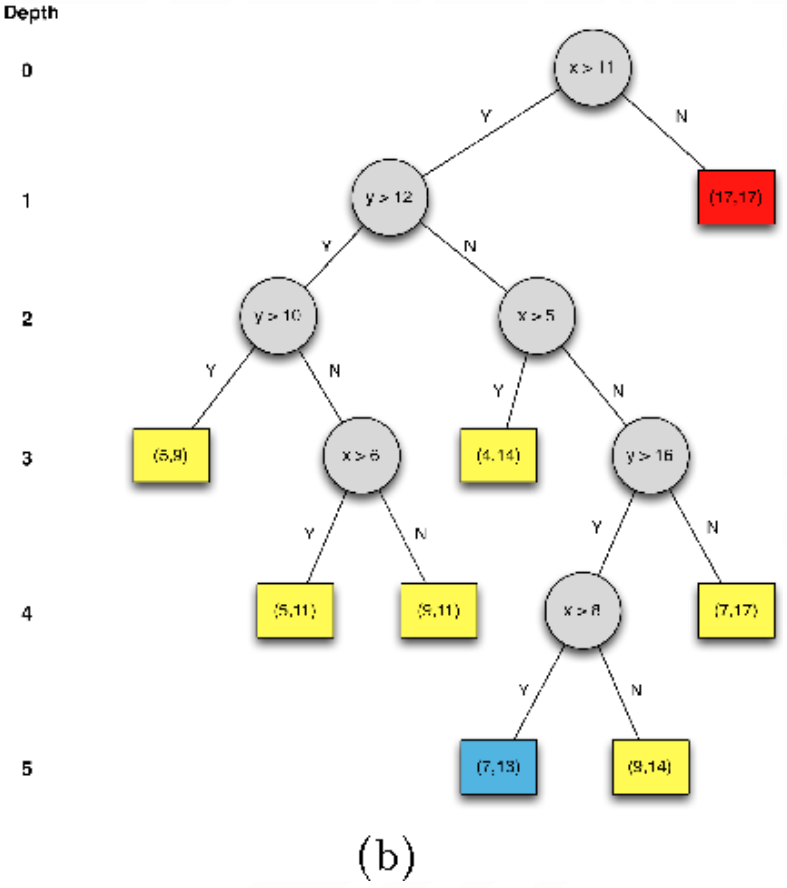
Идея: чем сильнее объект отличается от большинства, тем раньше он будет отделен от основной выборки случайными разбиениями => выбросы – объекты, которые оказались на небольшой глубине.



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.





ISOLATION FOREST

- Если объект единственный в листе, то его оценка аномальности в дереве – это глубина листа $\boldsymbol{h_n(x) = k}$.

- Оценка аномальности объекта в Isolation Forest:

$$\boldsymbol{a(x) = 2^{-\frac{a}{b}}},$$

где $a = \frac{1}{N}\sum h_n(x)$ – средняя глубина, где N – число деревьев в лесе,

$b = c(l)$ – средняя длина пути, посчитанная по всем объектам и всем деревьям в лесе, построенном по выборке размера l.

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.

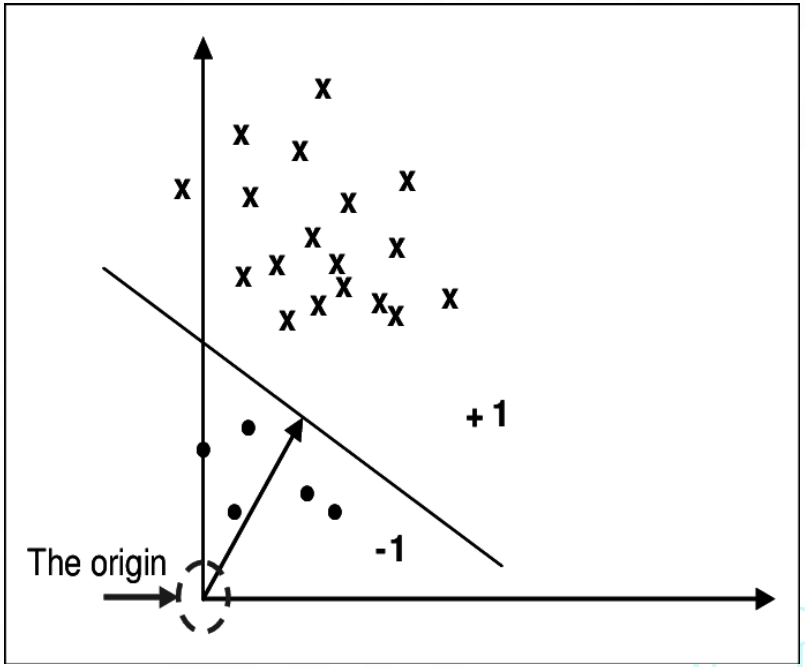
ONE-CLASS SVM

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

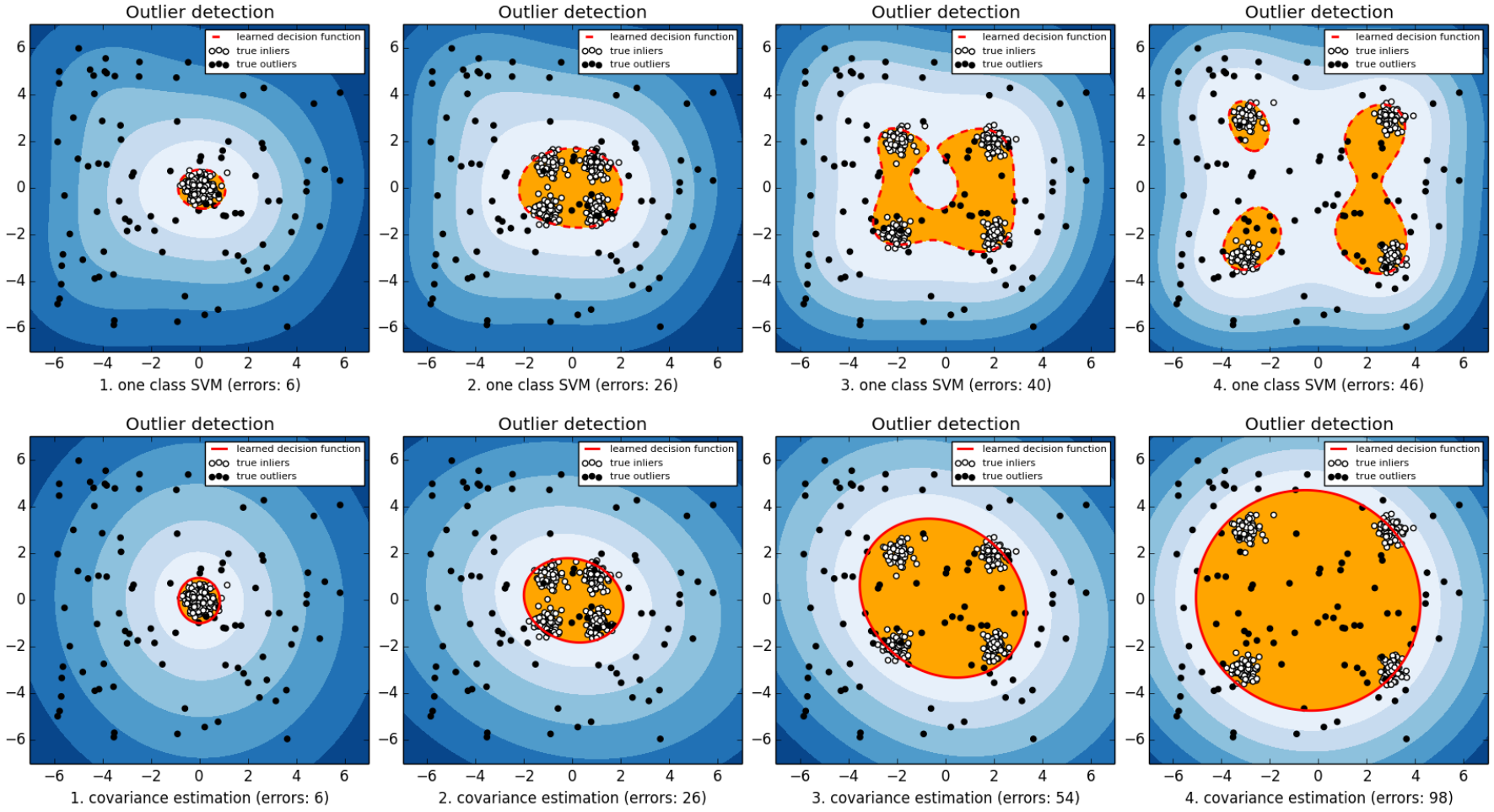
- $a(x)$ отделяет как можно больше объектов выборки от нуля: $a(x) = +1$ на области как можно меньшего объема, содержащей как можно больше объектов выборки
- имеет большой отступ от 0.

Тогда объекты с $a(x) = -1$

– это аномалии.

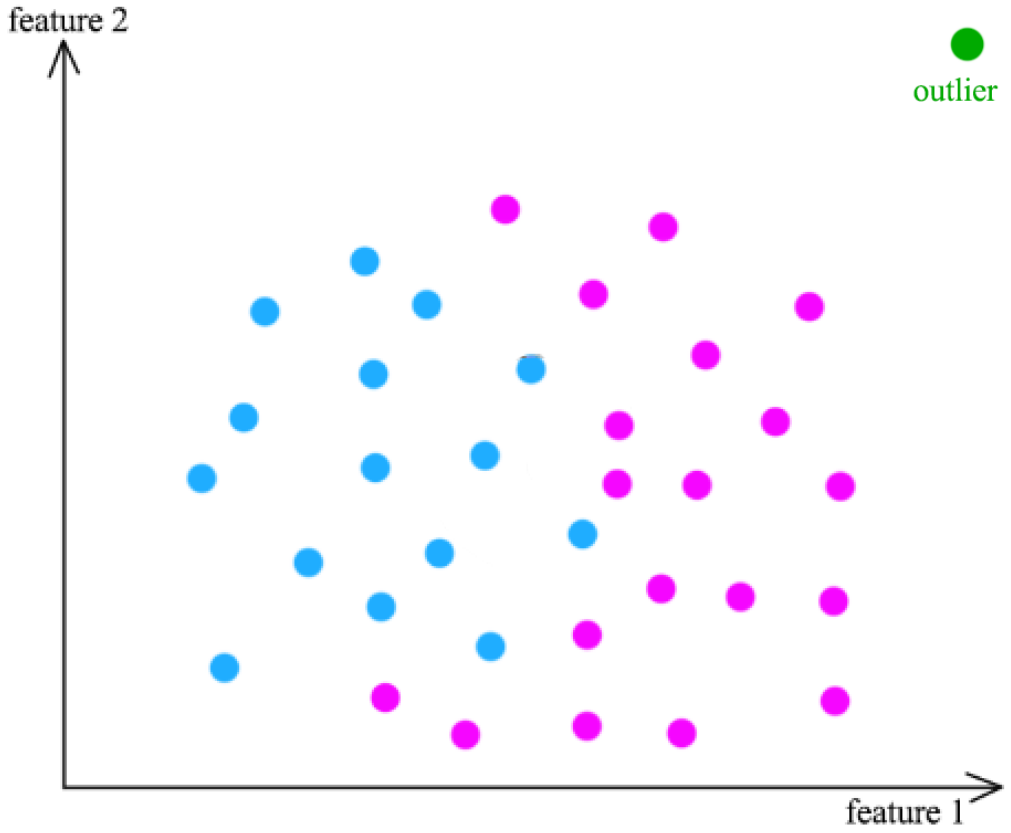


ONE-CLASS SVM С RBF-ЯДРОМ



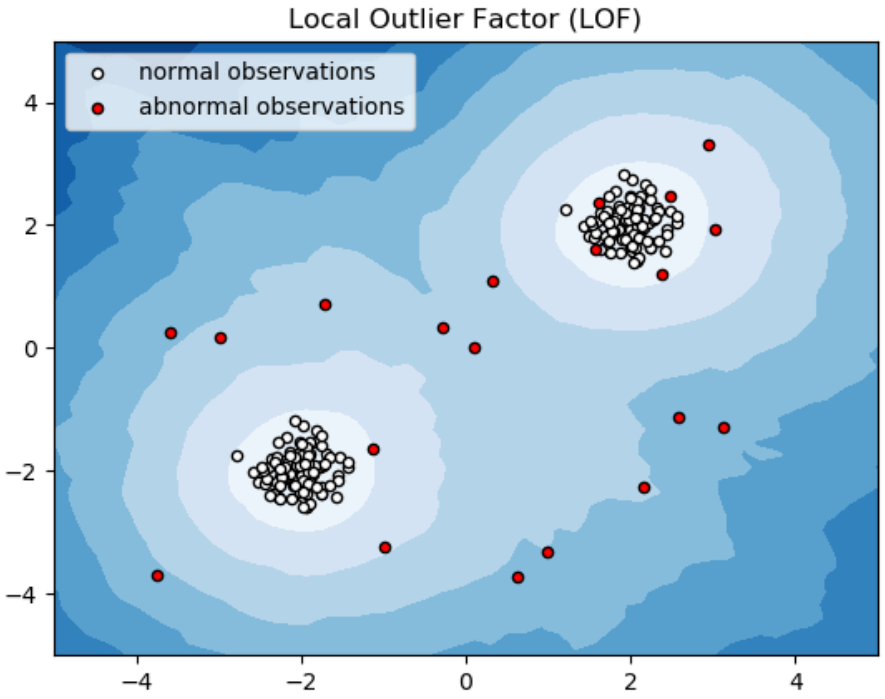
ПОИСК ВЫБРОСОВ С ПОМОЩЬЮ KNN

- 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.



ССЫЛКИ

- https://dyakonov.org/2017/04/19/поиск-аномалий-anomaly-detection/
- https://scikit-learn.org/stable/modules/outlier_detection.html
- https://github.com/yzhao062/pyod