



Лекция 13

Поиск аномалий.

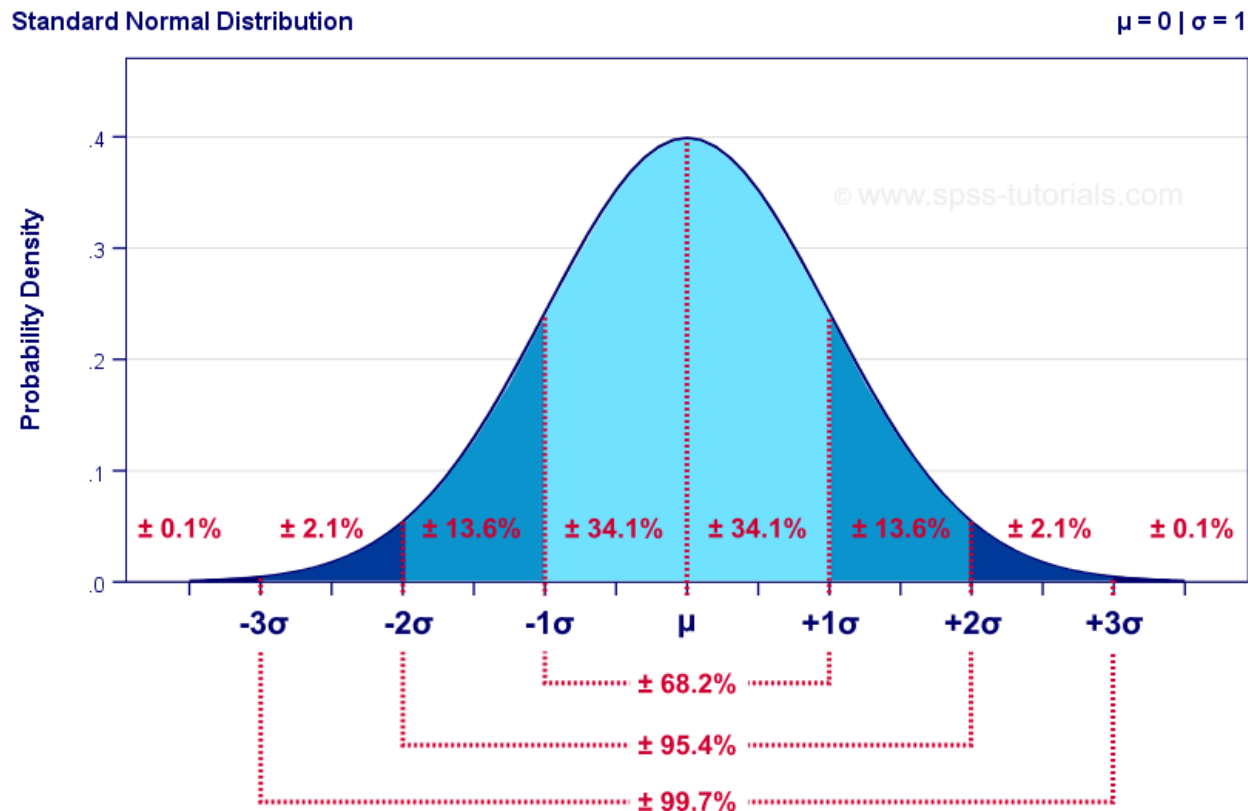
Кантонистова Е.О.

ВШЭ, 2021

Z-SCORE

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

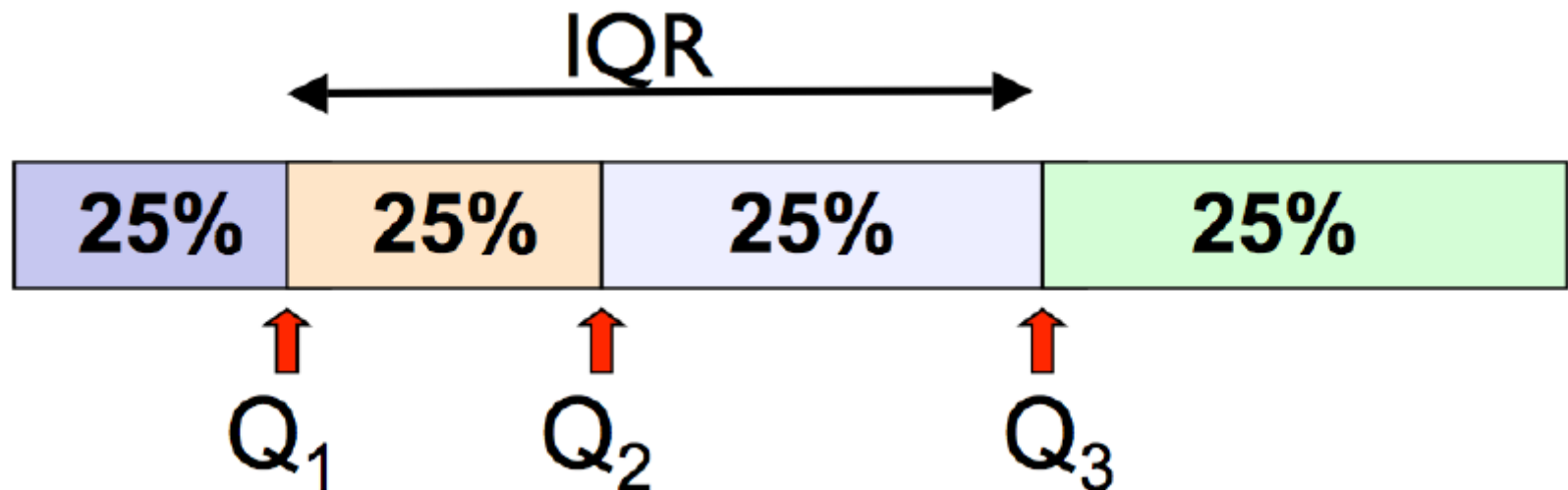
Точки, не попавшие в этот диапазон, можно считать выбросами.



НАХОЖДЕНИЕ ВЫБРОСОВ В ДАННЫХ

Пусть Q_1 – первая (25%) квартиль распределения,
 Q_3 – третья (75%) квартиль распределения.

- Величина $IQR = Q_3 - Q_1$ называется *интерквартильным размахом*.



НАХОЖДЕНИЕ ВЫБРОСОВ В ДАННЫХ

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

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

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

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

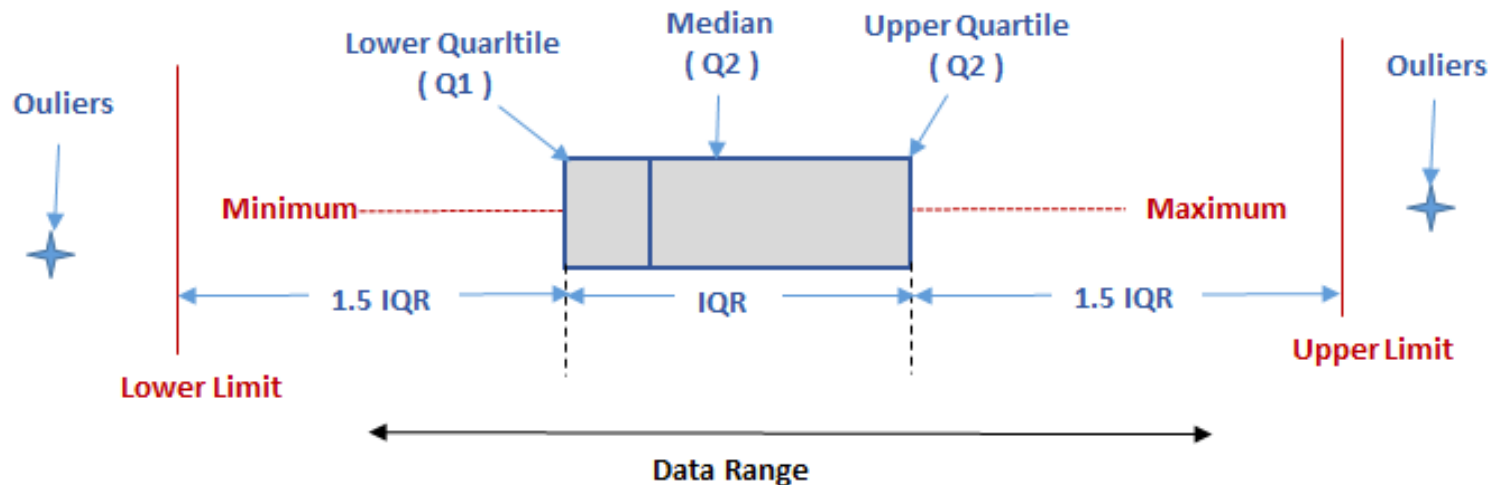


ДИАГРАММА ЯЩИК С УСАМИ (BOX-AND- WHISKER PLOT)

- Ящик с усами – это диаграмма, визуализирующая основные характеристики данных.



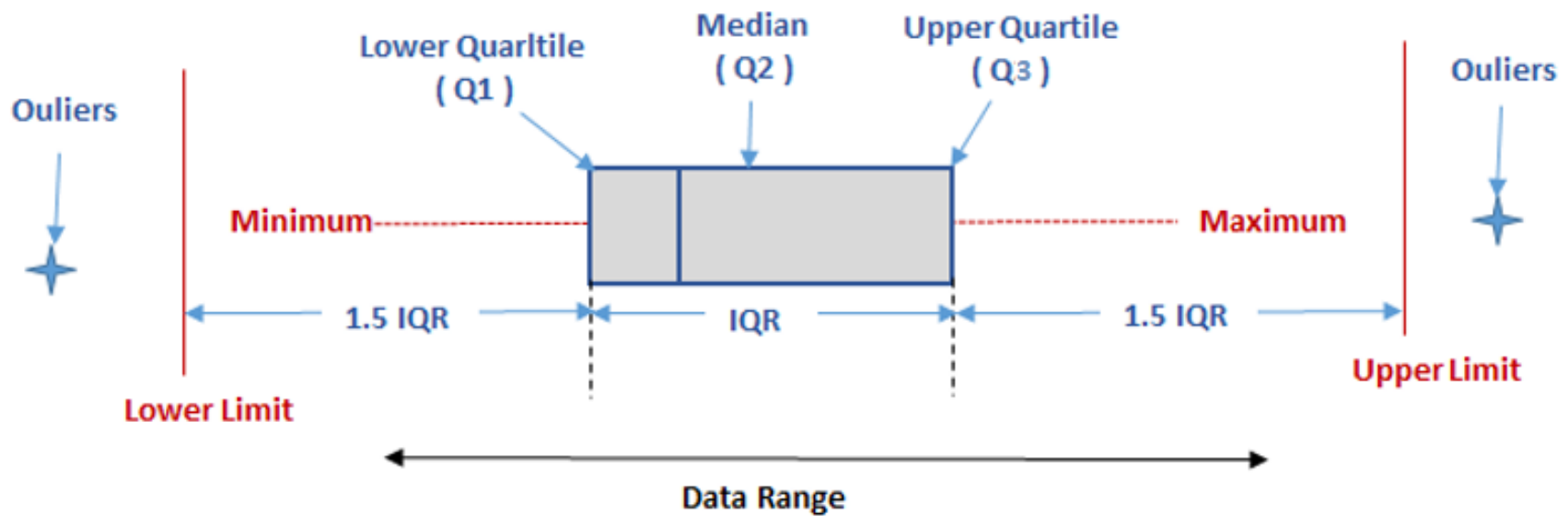
ДИАГРАММА ЯЩИК С УСАМИ (BOX-AND-WHISKER PLOT)

Ящик с усами – это диаграмма, визуализирующая основные характеристики данных.

Она состоит из:

- 1) Медианы** (“центрального” значения распределения)
- 2) Первой и третьей квартилей:** $Q1$ и $Q3$.
- 3) Минимума и максимума**
- 4)левой и правой границ, выйдя за которые точки считаются выбросами.**

Диаграмма ящик с усами (Box-and-Whisker plot)

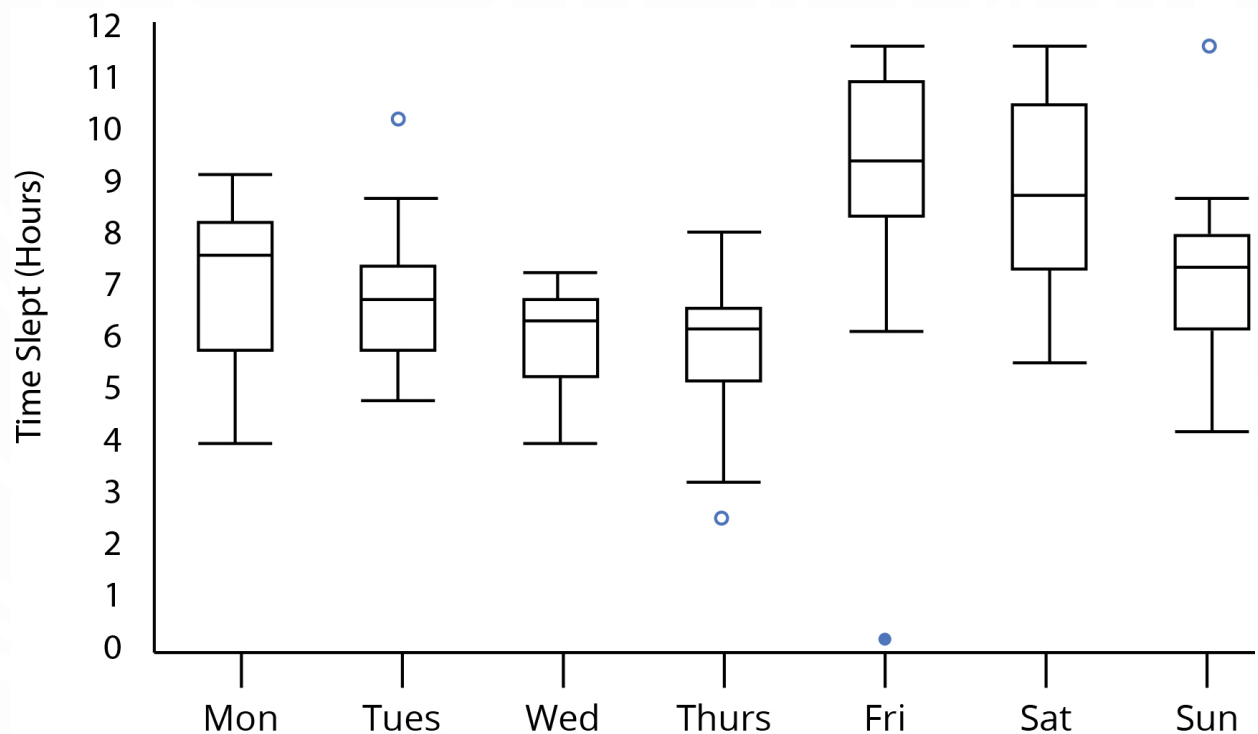


<http://www.whatissixsigma.net/box-plot-diagram-to-identify-outliers/>

ПРИМЕР: ДИАГРАММА ЯЩИК С УСАМИ

Очень легко видеть, как
распределяется
продолжительность сна в
зависимости от дня недели.

Хорошо видны выбросы и
общие тенденции.



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.

Grow a random decision tree until each instance is in its own leaf

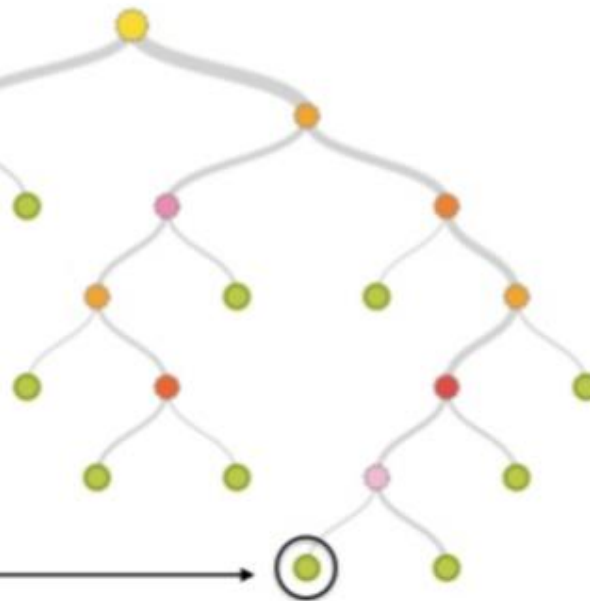
“easy” to isolate →



Depth

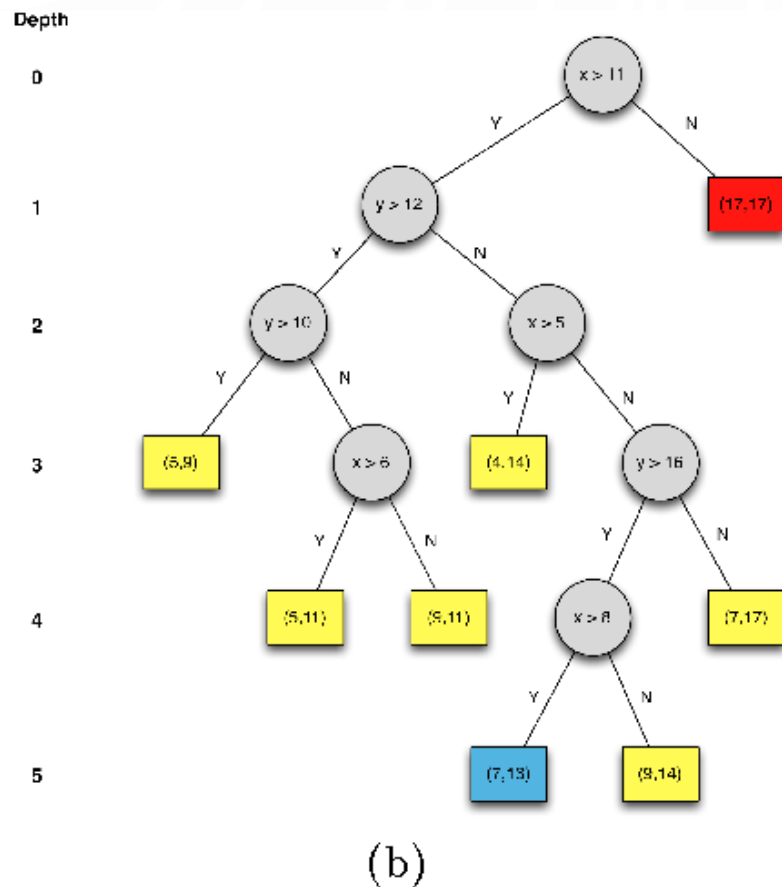
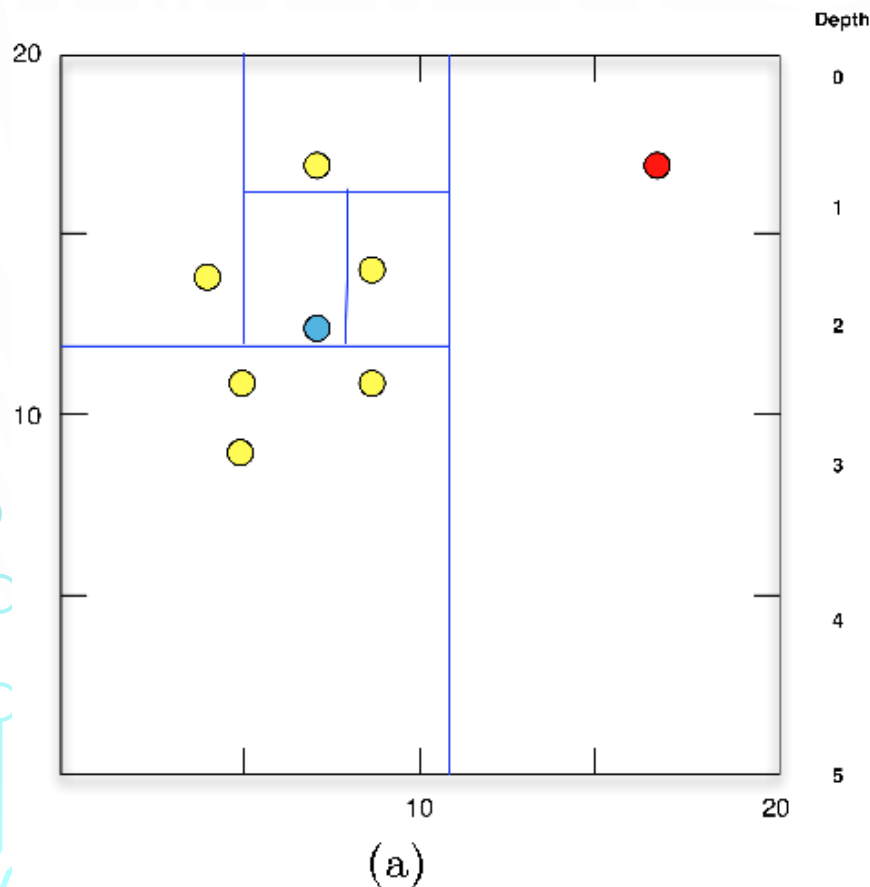
“hard” to isolate →

Now repeat the process several times and use average Depth to compute anomaly score: 0 (similar) -> 1 (dissimilar)



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.

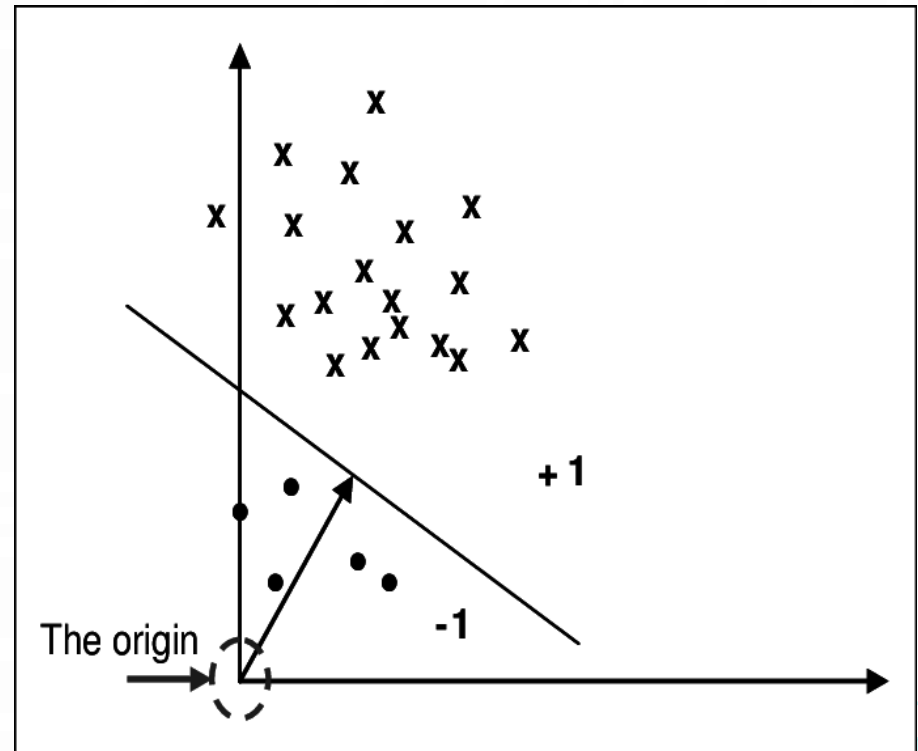


ONE-CLASS SVM

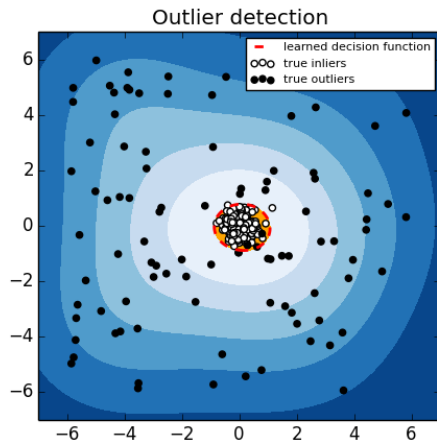
Метод строит линейную функцию $a(x) = \text{sign}(w, x)$ так, чтобы она отделяла выборку от начала координат с максимальным отступом, а именно:

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

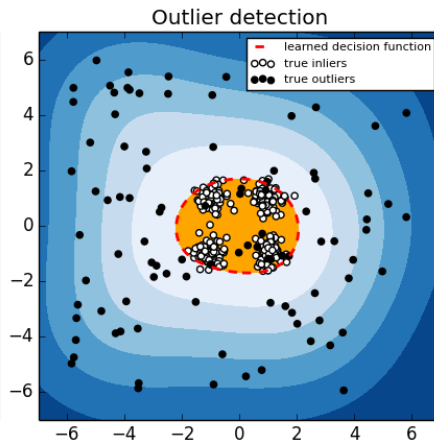
Тогда объекты с $a(x) = -1$
— это аномалии.



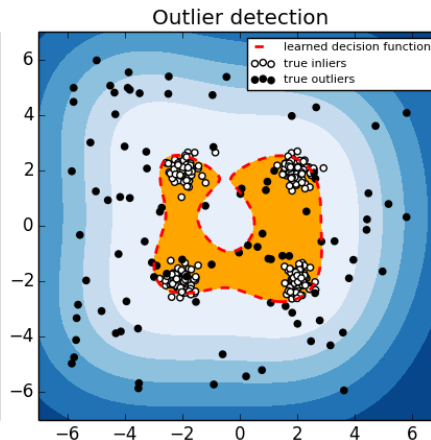
ONE-CLASS SVM С RBF-ЯДРОМ



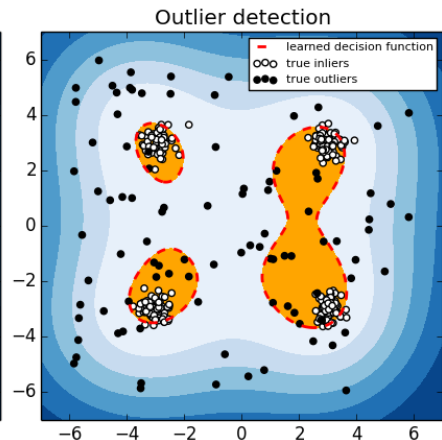
1. one class SVM (errors: 6)



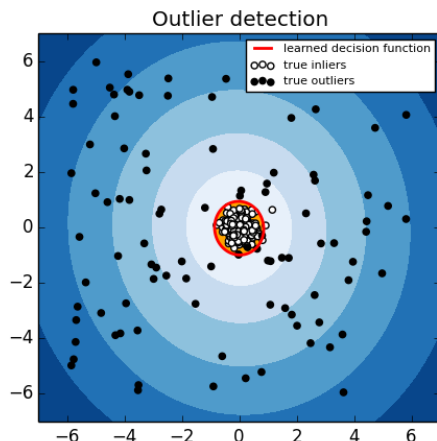
2. one class SVM (errors: 26)



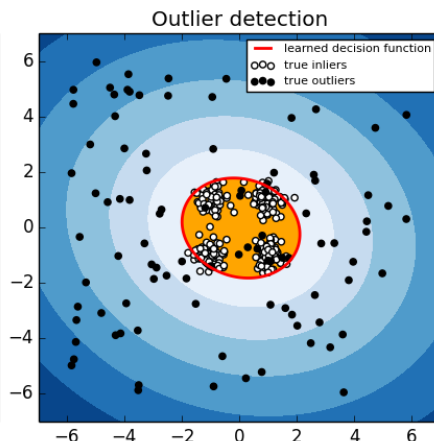
3. one class SVM (errors: 40)



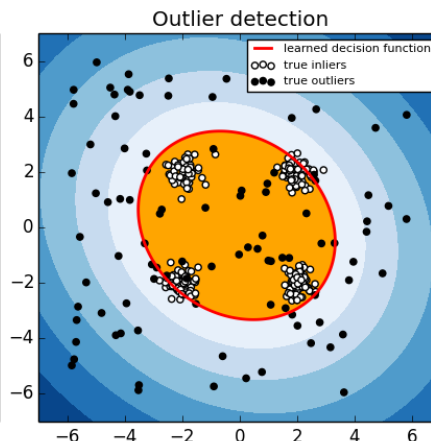
4. one class SVM (errors: 46)



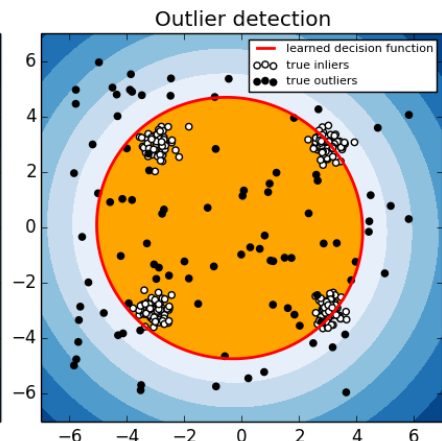
1. covariance estimation (errors: 6)



2. covariance estimation (errors: 26)



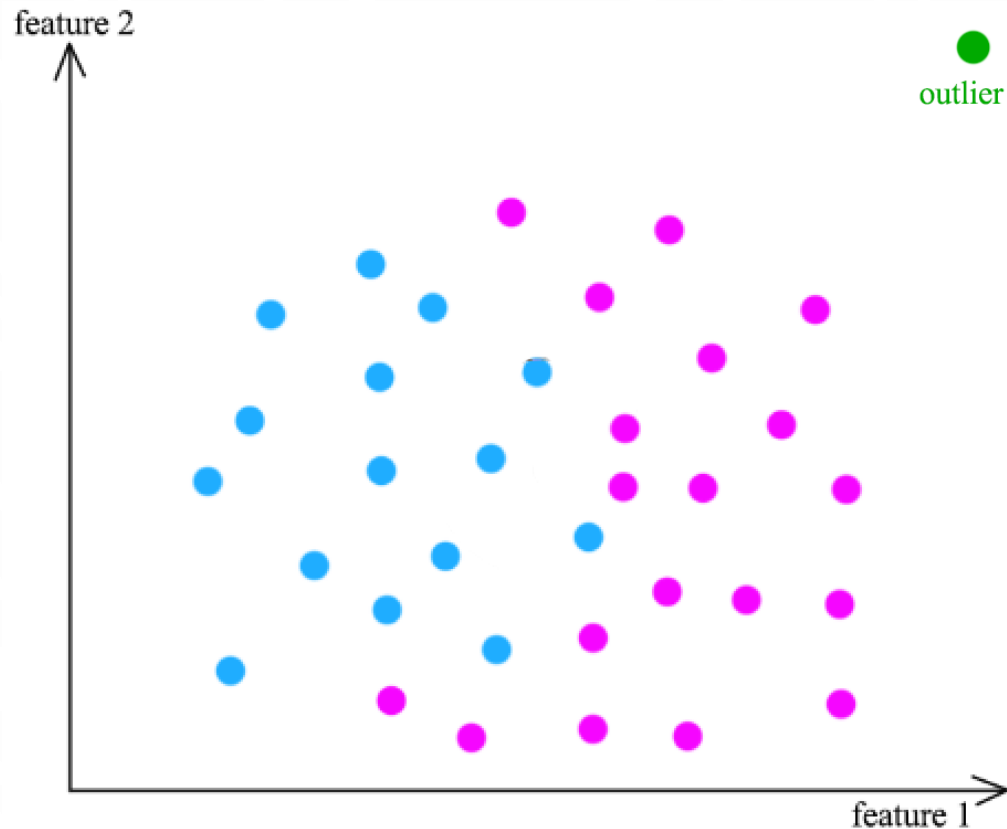
3. covariance estimation (errors: 54)



4. covariance estimation (errors: 98)

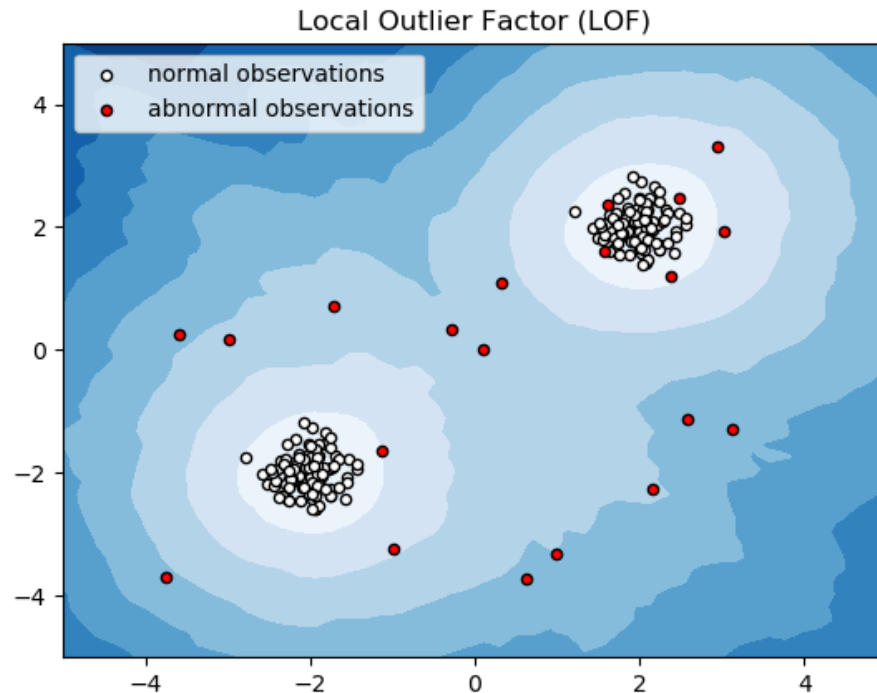
ПОИСК ВЫБРОСОВ С ПОМОЩЬЮ KNN

- 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>