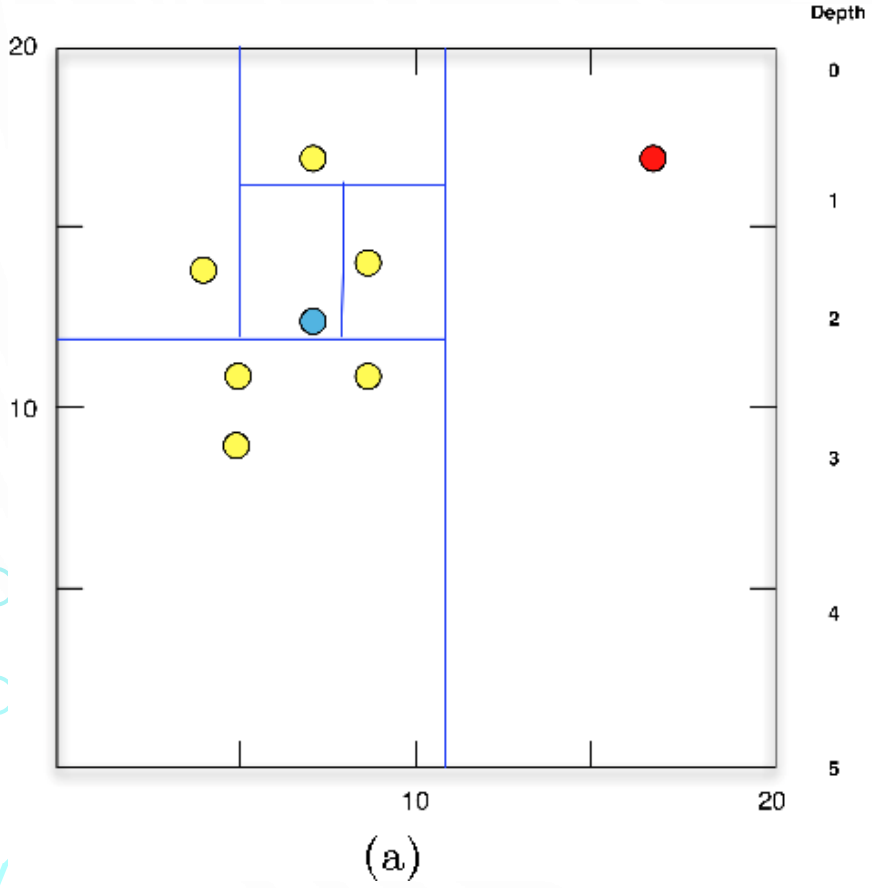
Лекция 10
Поиск аномалий.

Кантонистова Е.О.

ВШЭ, 2019

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.



ПОИСК АНОМАЛИЙ С ПОМОЩЬЮ МОДЕЛЕЙ ML

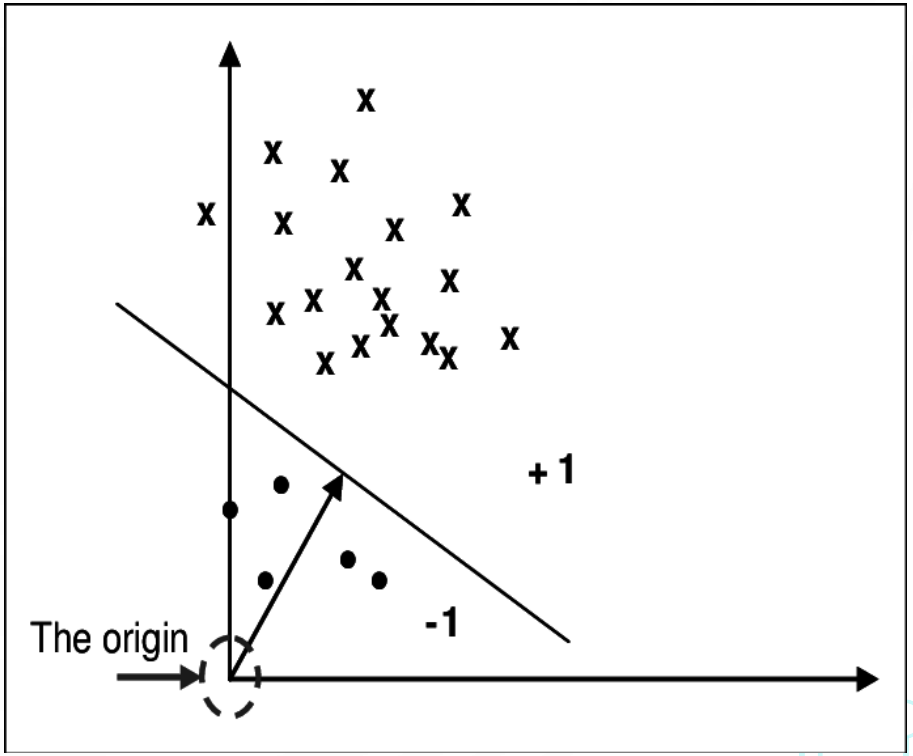
Идея: можно настроить модель машинного обучения так, чтобы на нормальных объектах она принимала значения, близкие к нулю (или, например, положительные значения). Тогда если прогноз на объекте сильно отличается от прогноза на обучающей выборке, то такой объект можно считать аномальным.

ONE-CLASS SVM

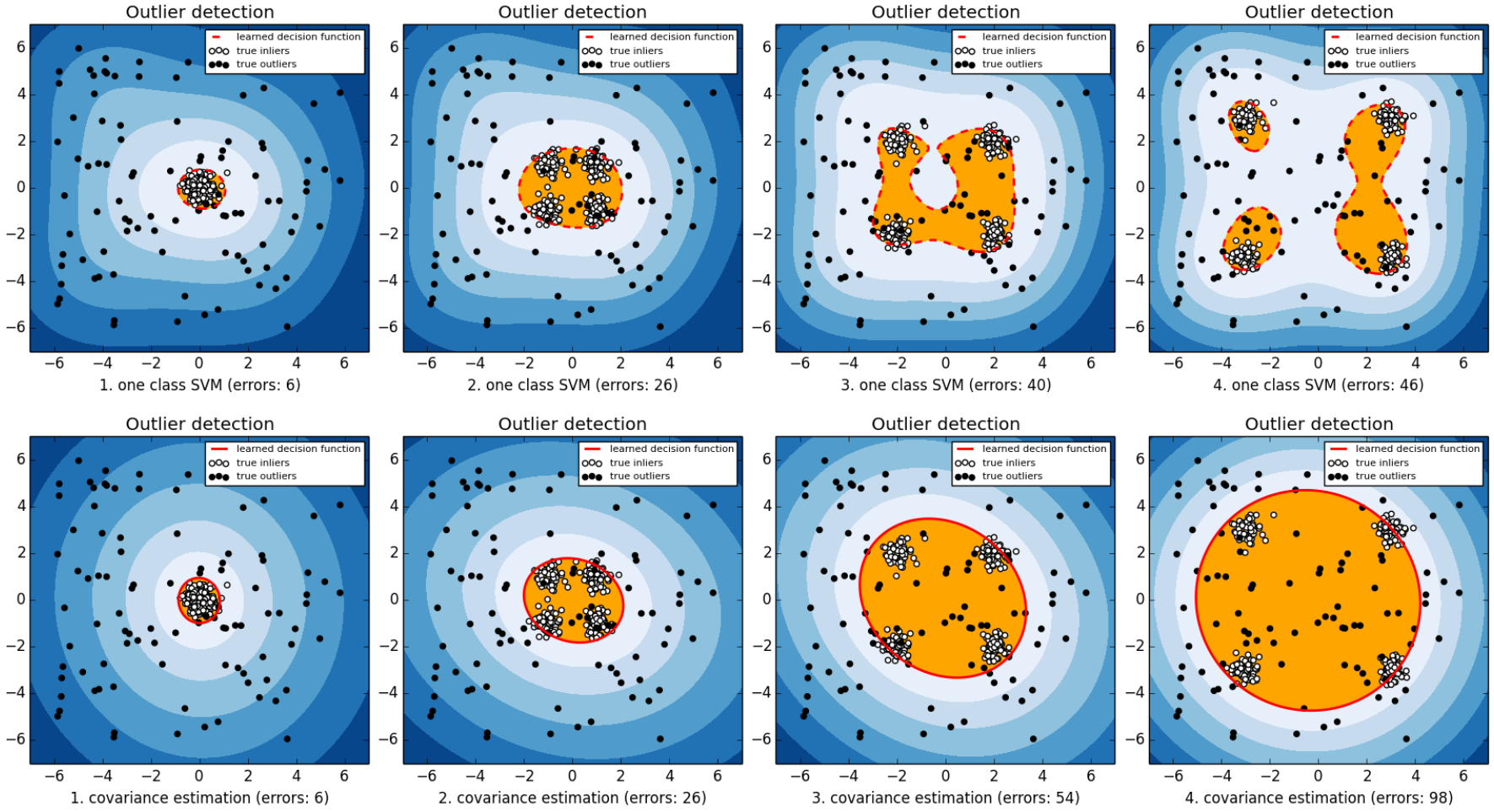
Метод строит линейную функцию $a(x) = sign(w, x)$ так, чтобы она отделяла выборку от начала координат с максимальным отступом, а именно:

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

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

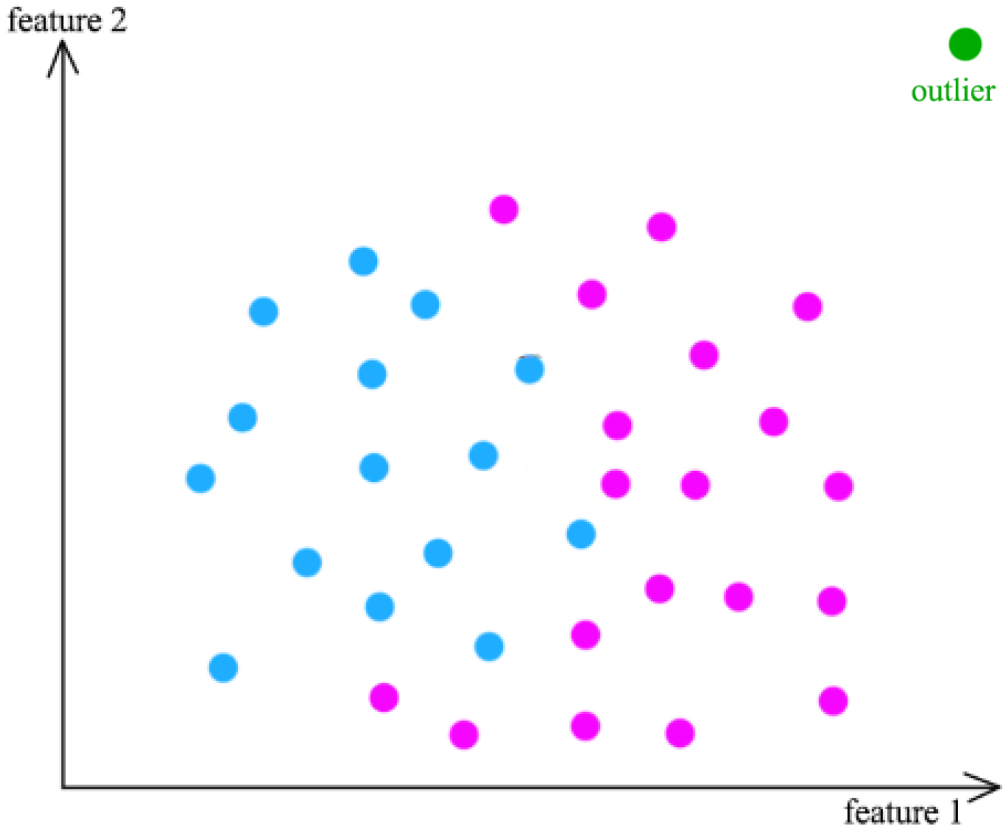


ONE-CLASS SVM С RBF-ЯДРОМ



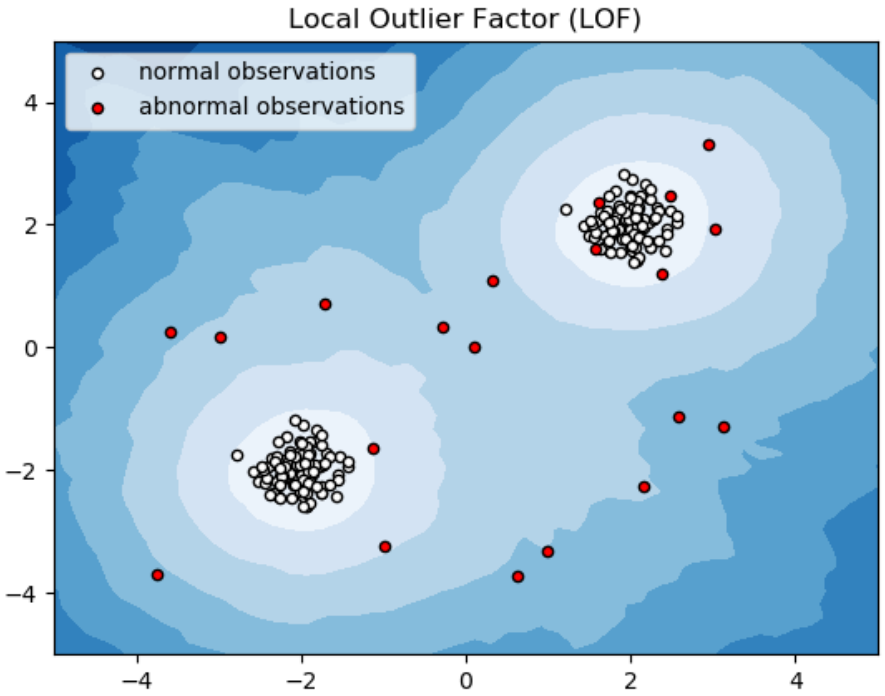
ПОИСК ВЫБРОСОВ С ПОМОЩЬЮ KNN

- 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