Лекция 4. Устранение проблемы наличия выбросов и аномалий в данных

Практикум по программированию, 5 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Что такое выбросы и аномалии в данных?

Выбросы – это значения данного столбца, которые не похожи на подавляющее большинство значений

Пример 1. Выбросы в данных

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	45
79.6	b	4	34
45664.8	b	6	64
67.08	a	6	23
86.7	b	5	2348

Аномалии – это объекты (строки), которые не похожи на подавляющее большинство объектов

Пример 2. Аномальные объекты в данных

num_feature1	cat_feature1	num_feature2	target
56.09	a	7	0
79.6	a	4	1
64.6	a	7	1
778976	b	536545	1
60.6	a	6	0

Инструменты по борьбе с выбросами

Данные подходы пытаются найти пороговые величины, которые заключают большинство схожих значений внутри признака

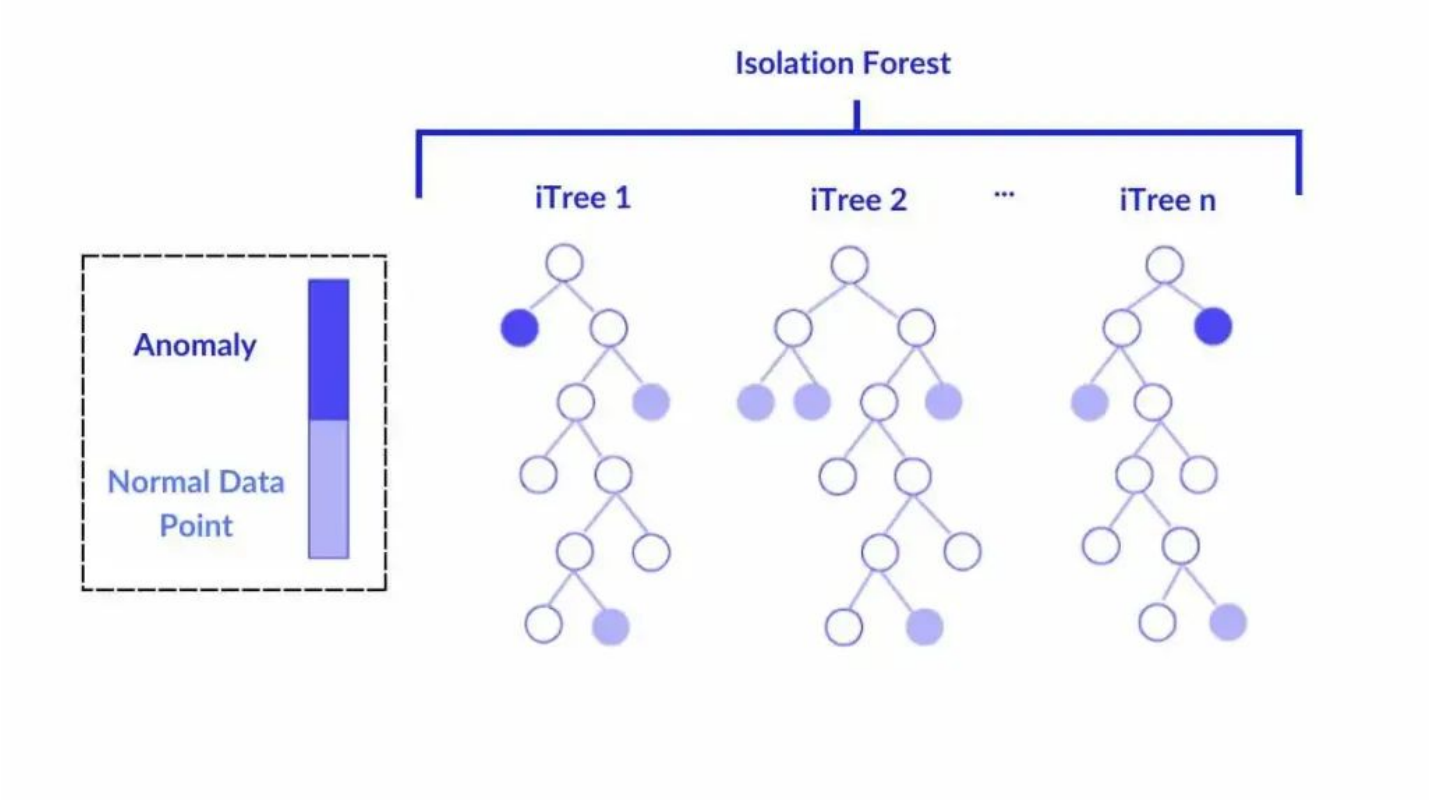
1. Анализ распределений
2. IQR
3. Z-score

Инструменты по борьбе с аномалиями

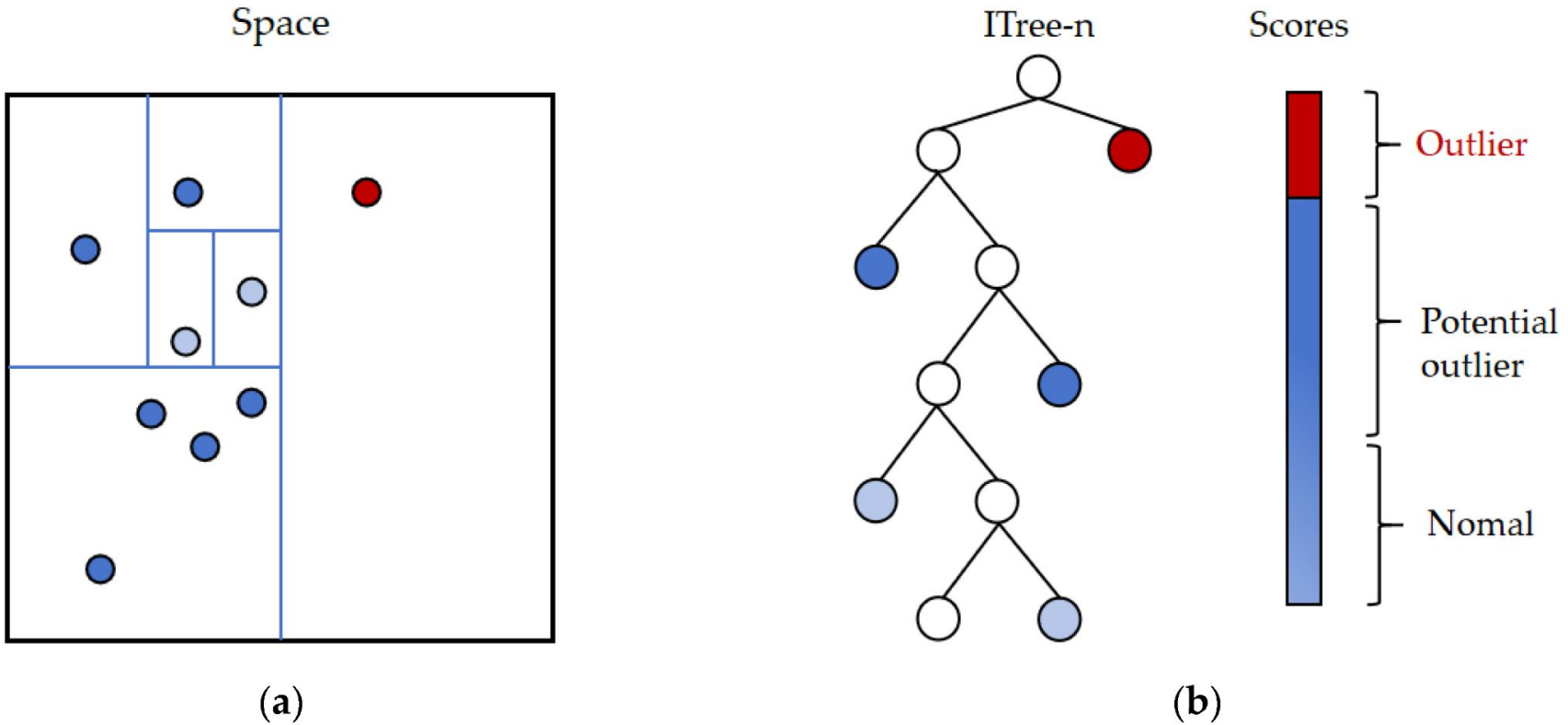
Данные подходы (обучение без учителя) пытаются анализировать различные аспекты пространственной информации и тем самым найти аномальные объекты

1. Isolation Forest
2. One-class SVM
3. DBSCAN

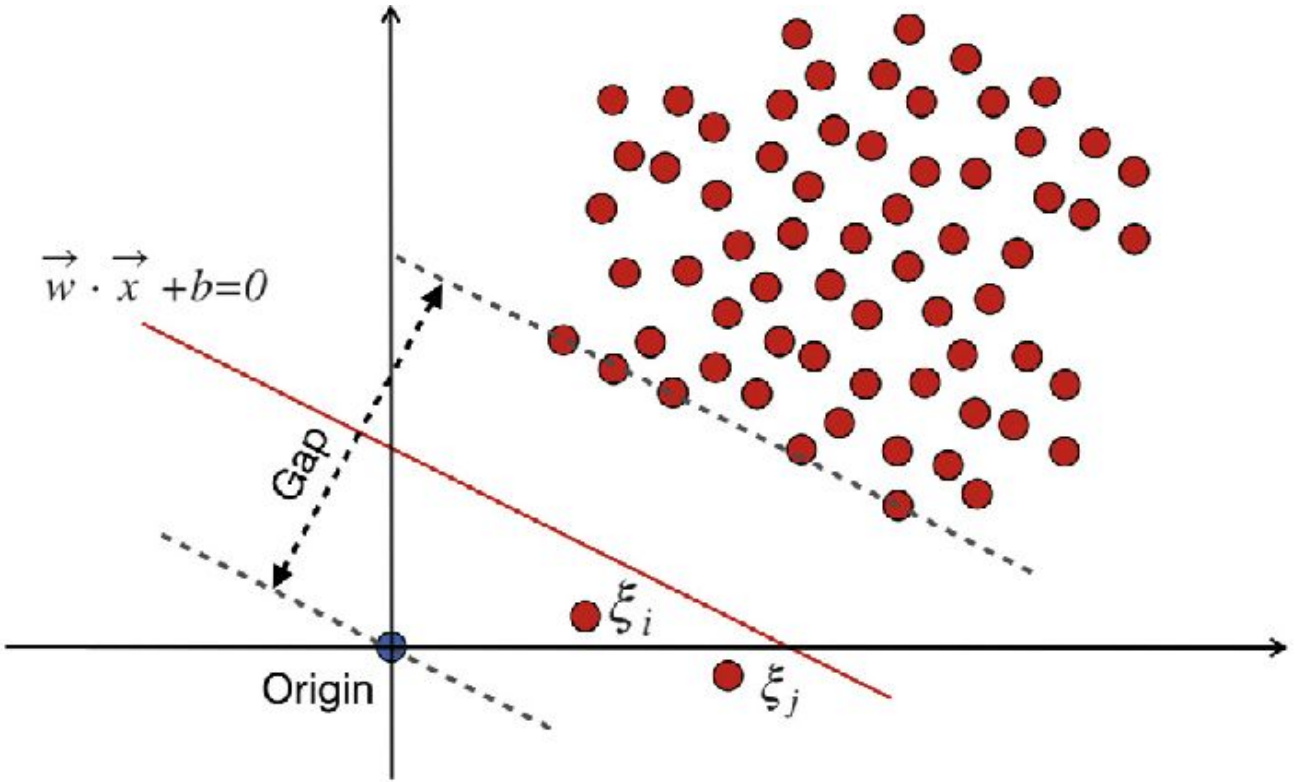
Isolation Forest



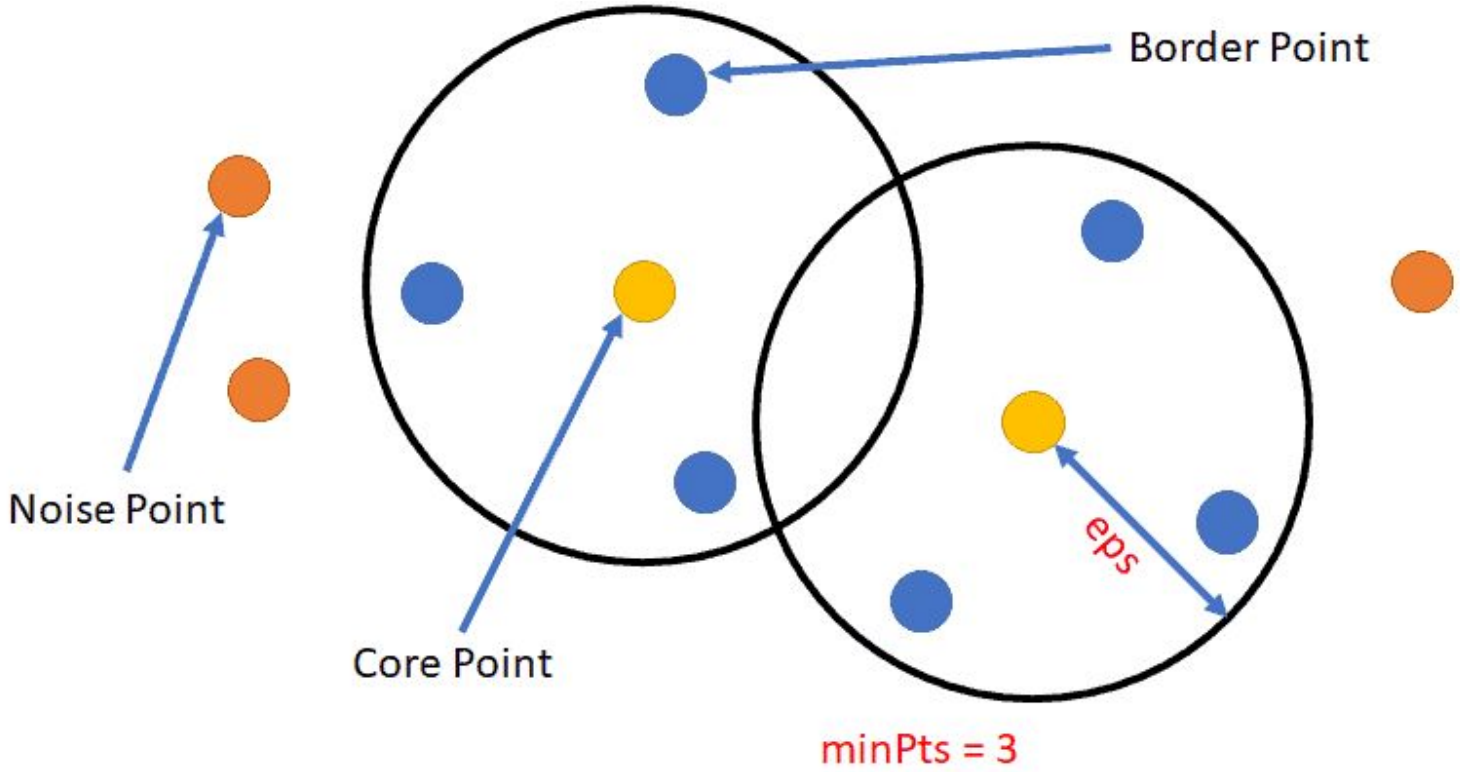
Isolation Forest



Isolation Forest



Isolation Forest



Какой подход выбрать?

Советы по работе с аномалиями и выбросами

1. Всегда начинайте с построения гистограмм, функций плотности и диаграмм "Ящик с усами": они покажут наличие выбросов внутри отдельных признаков
2. Если видны выбросы, то применяйте IQR / Z-score (а лучше использовать и то, и то, а потом посмотреть пересечение результатов). Помните, что выбросы необходимо **интерпретировать (!)**
3. После работы с выбросами можно посмотреть аномальные объекты (также лучше применять несколько подходов и сравнивать результаты). А потом попытаться их **интерпретировать (!)**
4. Всегда можно визуализировать аномалии в двумерном пространстве
5. Работать с аномалиями и выбросами очень сложно, **без экспертной оценки и понимания предметной области удалять строки нельзя (!)**

Виды кодировщиков категориальных признаков

Метод	Плюсы	Минусы
One-Hot	Простота, интерпретируемость	Проклятие размерности для признаков с >10 категориями
Label	Простота, не увеличивает размерность	Произвольный порядок, плохо для линейных моделей
Target	Учитывает связь с целью, сохраняет размерность	Риск переобучения, чувствительность к настройкам