Поиск аномалий

Смоляков Дмитрий

Skoltech

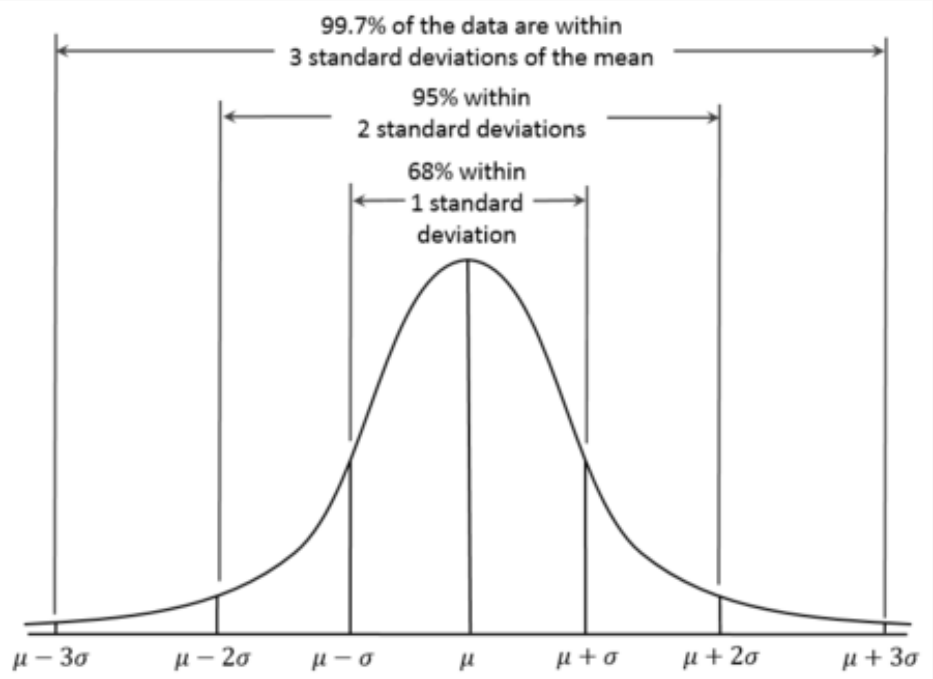
Содержание

Дополнительные материалы и презентация
`https://github.com/sklef/datastart`

Что такое аномалии?

Определение аномалии/выброса [Howkins, 1980]

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”



KDD-99

KDD-99 Dataset

- Симуляция 9 недель работы U.S. Airforce LAN
- Содержит примеры атак
- Более 7 миллионов записей
- 22 вида атак

KDD-99 Dataset

Можно выделить три группы признаков

TCP Dump	Экспертные признаки	Окно в 2 секунды
duration	login attempt	serror rate
protocol type	sudo attempt	same srv rate
flag	root login	diff srv rate
etc	etc	etc

KDD-99 Dataset

Подготовка признаков

- Численные значения нормировались
- Категориальных признаки кодировались One Hot Encoding

Финальный размер

- Порядка 7млн наблюдений
- Размерность 118

Задача

Задача

Находить атаки, имея на руках только информацию о нормальном функционировании

Решение

- Строим модель детектирования аномалий
- Аномальные наблюдения считаем атаками

Оценка качества

Precision/Recall

Precision = $\frac{TP}{TP+FP}$

Доля верных сигналов тревоги

Recall = $\frac{TP}{TP+FN}$

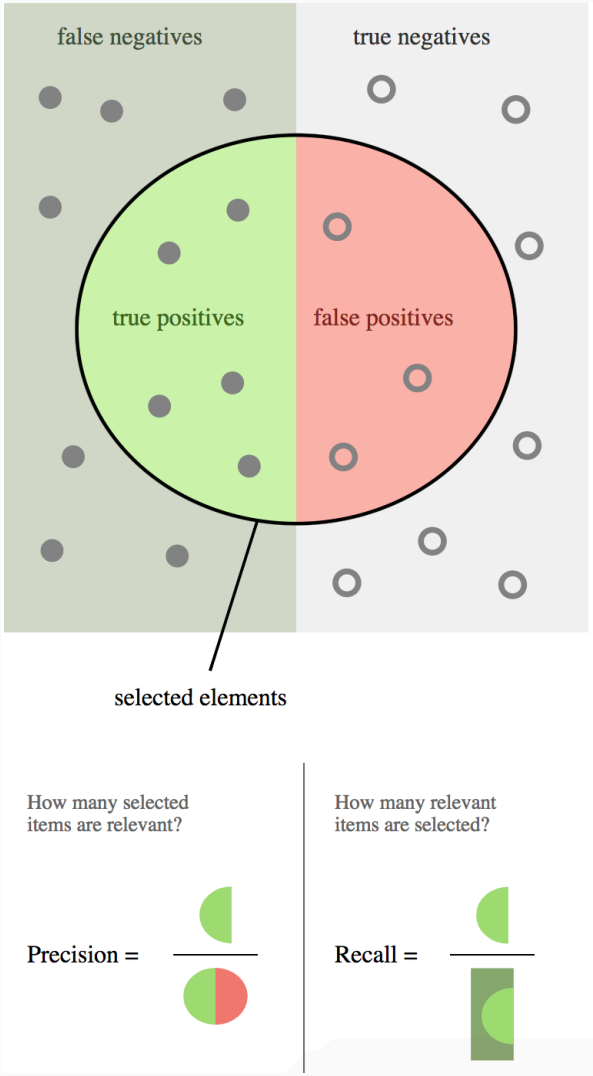
Доля найденных поломок

F1 = $2 \cdot \frac{precision \cdot recall}{precision+recall}$

Среднее гармоническое precision и recall

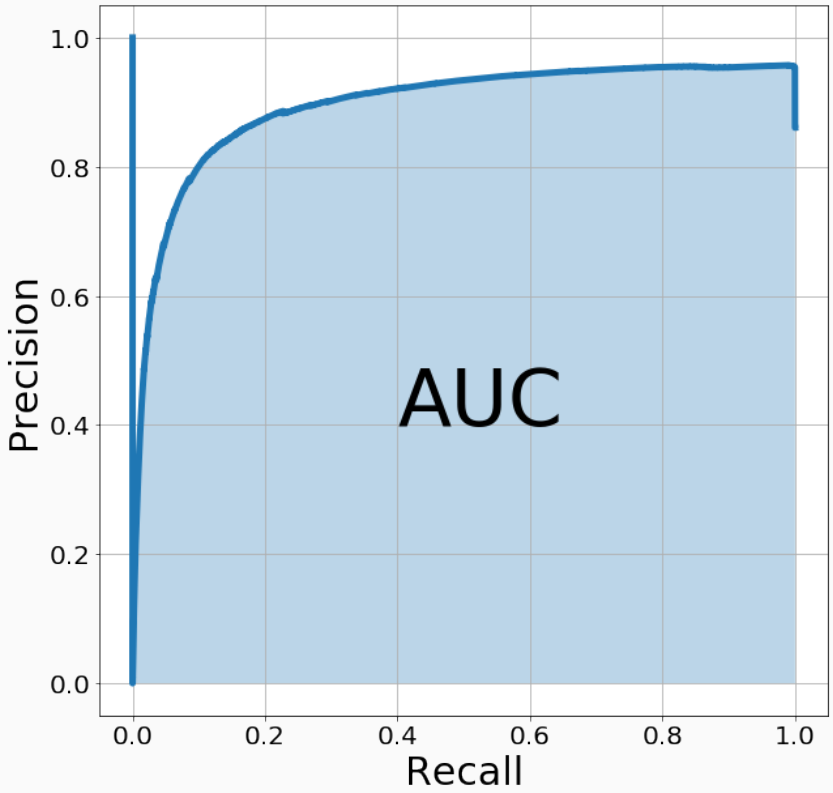
Fβ = $(1+\beta^2) \cdot \frac{precision \cdot recall}{\beta^2 precision+recall}$

Взвешенное среднее гармоническое precision и recall



Кривая Precision/Recall

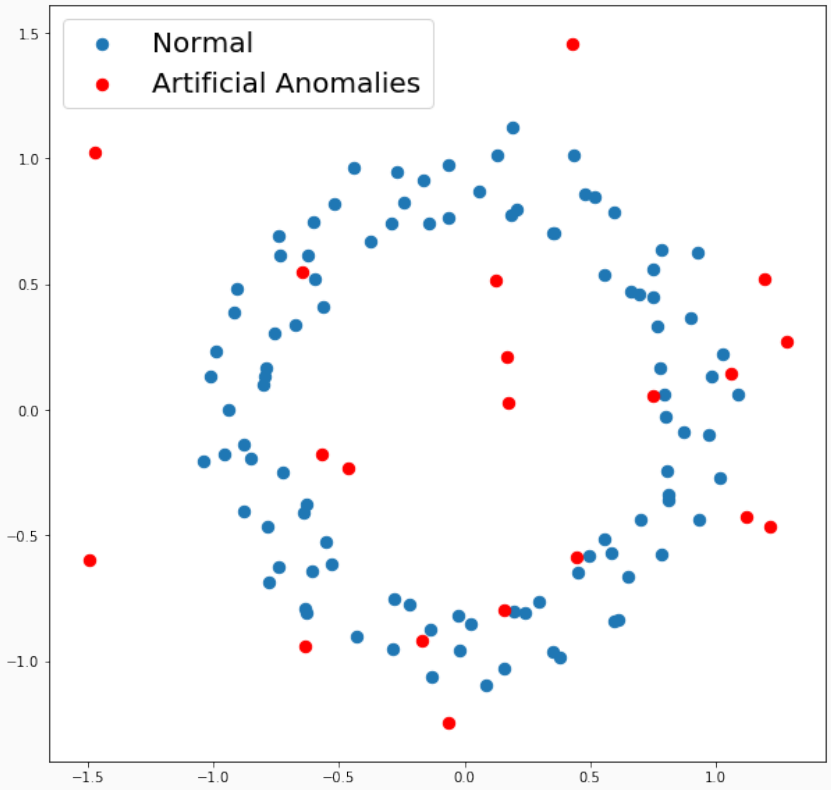
Если алгоритм детектирования аномалий позволяет выдавать степень уверенности, то в зависимости от порога отсечения можно получить разные значение precision и recall



Искусственная разметка

Если данных об аномалиях нет, их можно сгенерировать самому

- Априорные знания об аномалиях можно выразить через их распределение
- Если никаких сведений нет, можно воспользоваться равномерным распределением



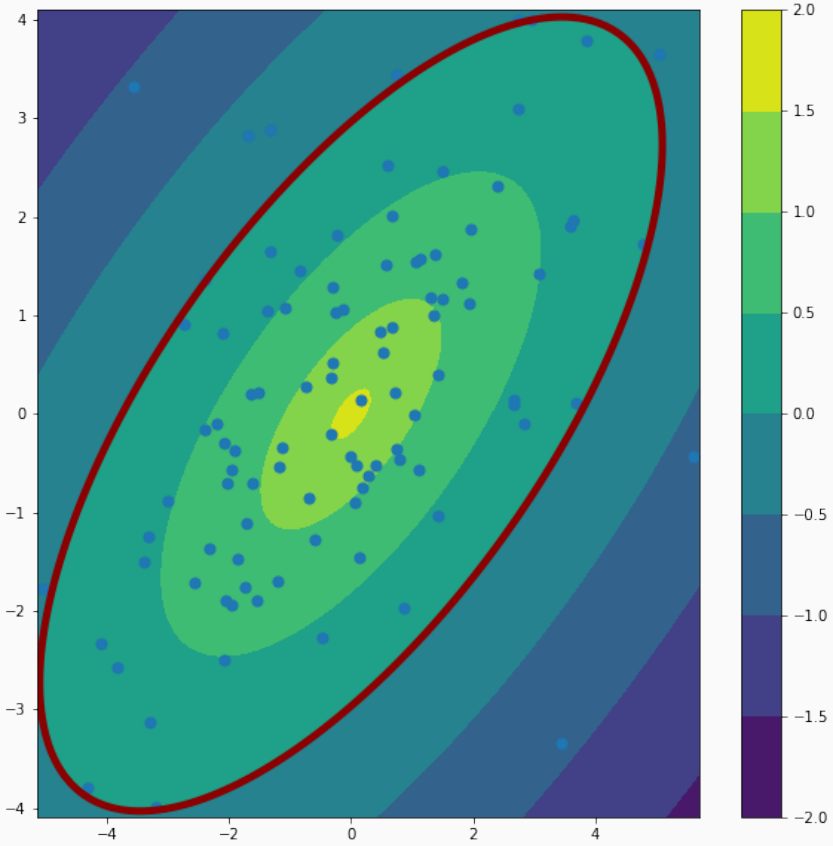
Методы детектирования аномалий

Elliptic Envelope

Считаем, что данные из нормального распределения

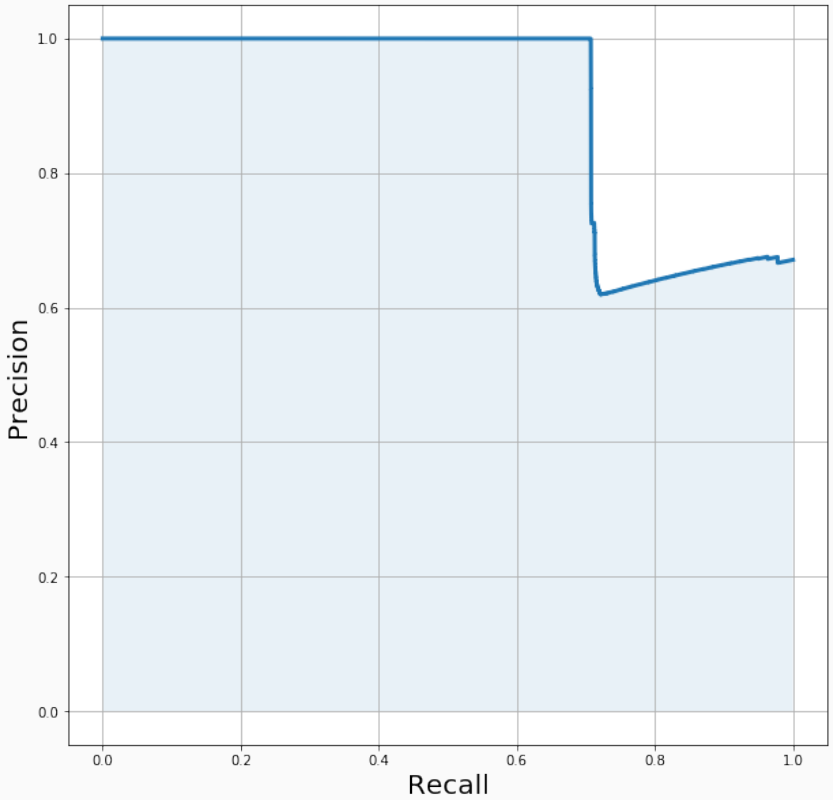
$$p(x|\mu,\Sigma) = \frac{\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)}{(2\pi)^{\frac{n}{2}}det(\Sigma)}$$

Используем робастную оценку ковариационной матрицы



Elliptic Envelope

Время обучения	1min 7s
Время на предсказание	1s
Precision/Recall AUC	0.9



Elliptic Envelope

Pros:

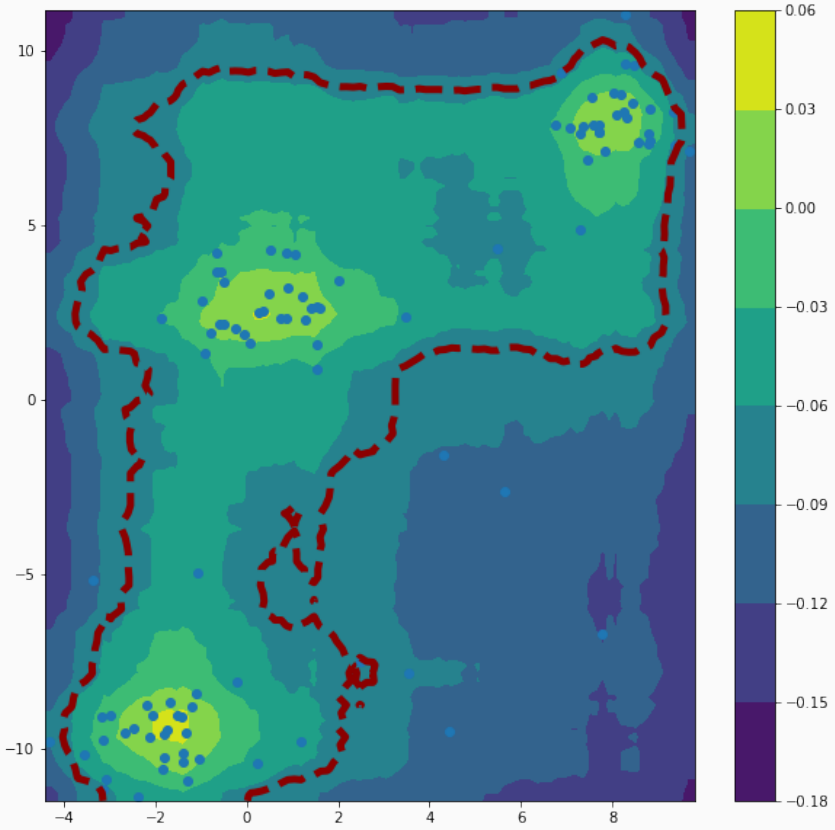
- Прост в использовании
- Легко интерпретировать

Cons:

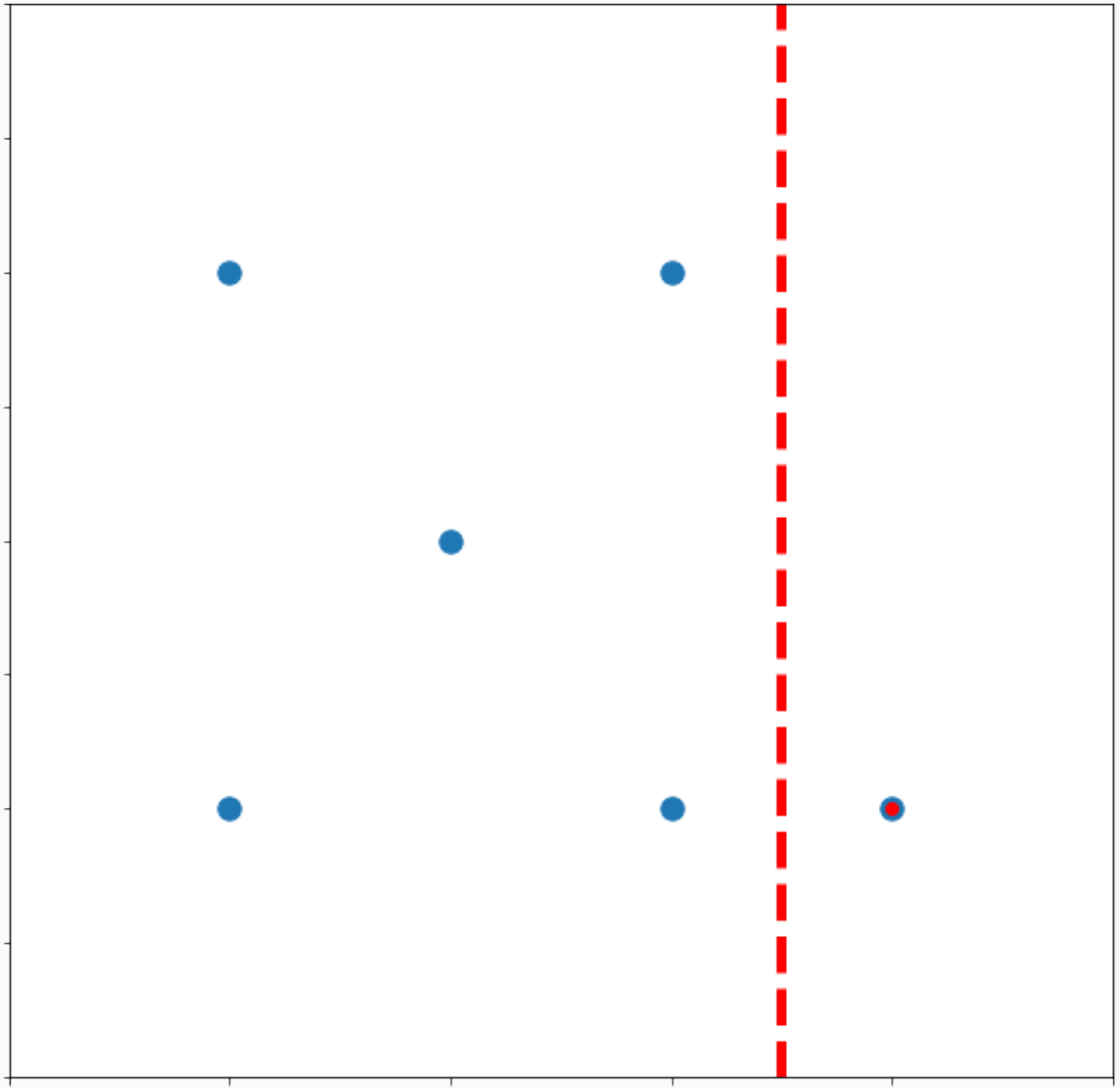
- Применим только для унимодальных распределений
- Плохо работает с коллинеарными данными

Isolation Forest

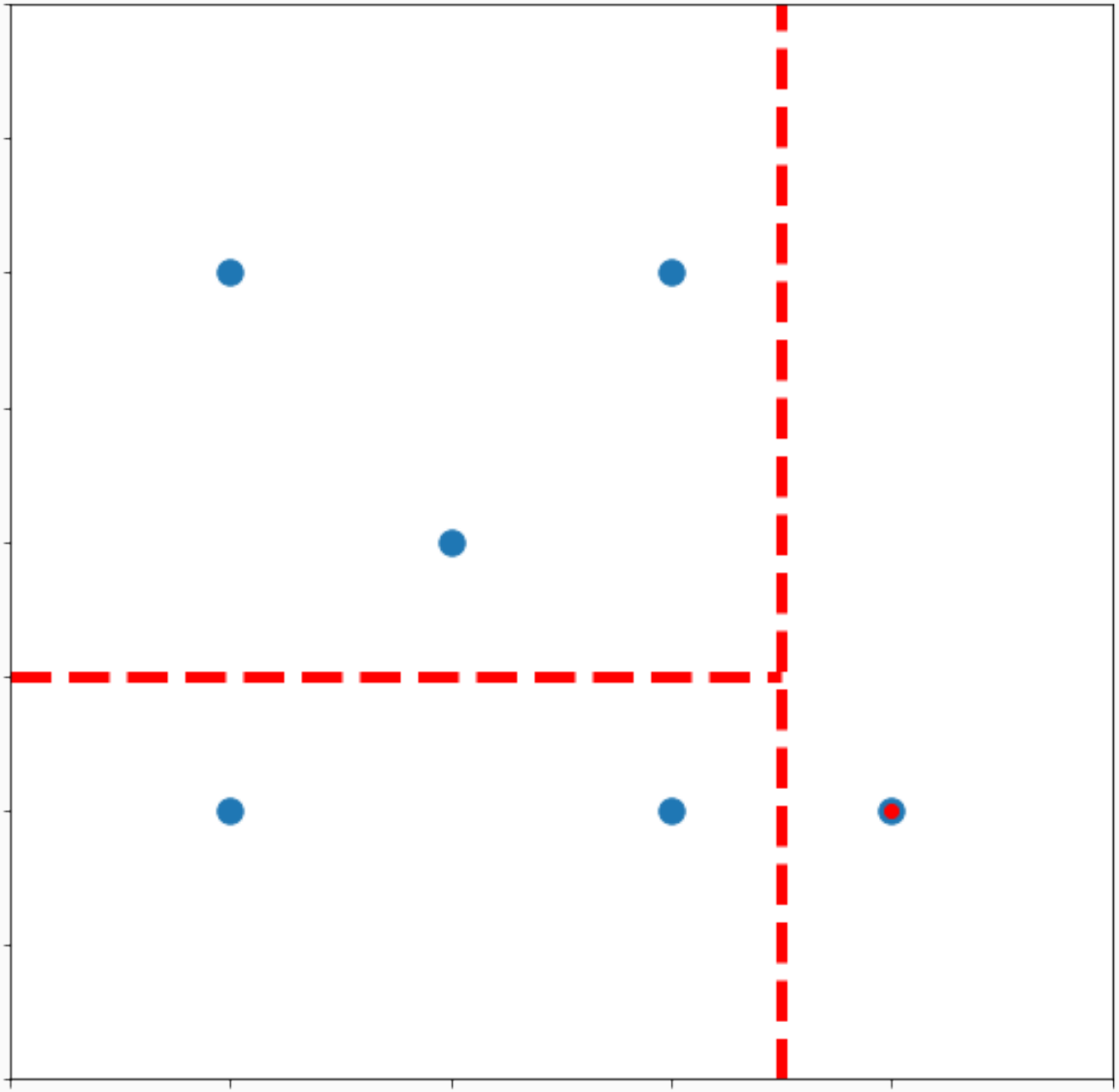
На каждой итерации производим случайное разбиение по случайному признаку. Чем меньше требуется разбиений, чтобы изолировать наблюдение тем более оно аномально



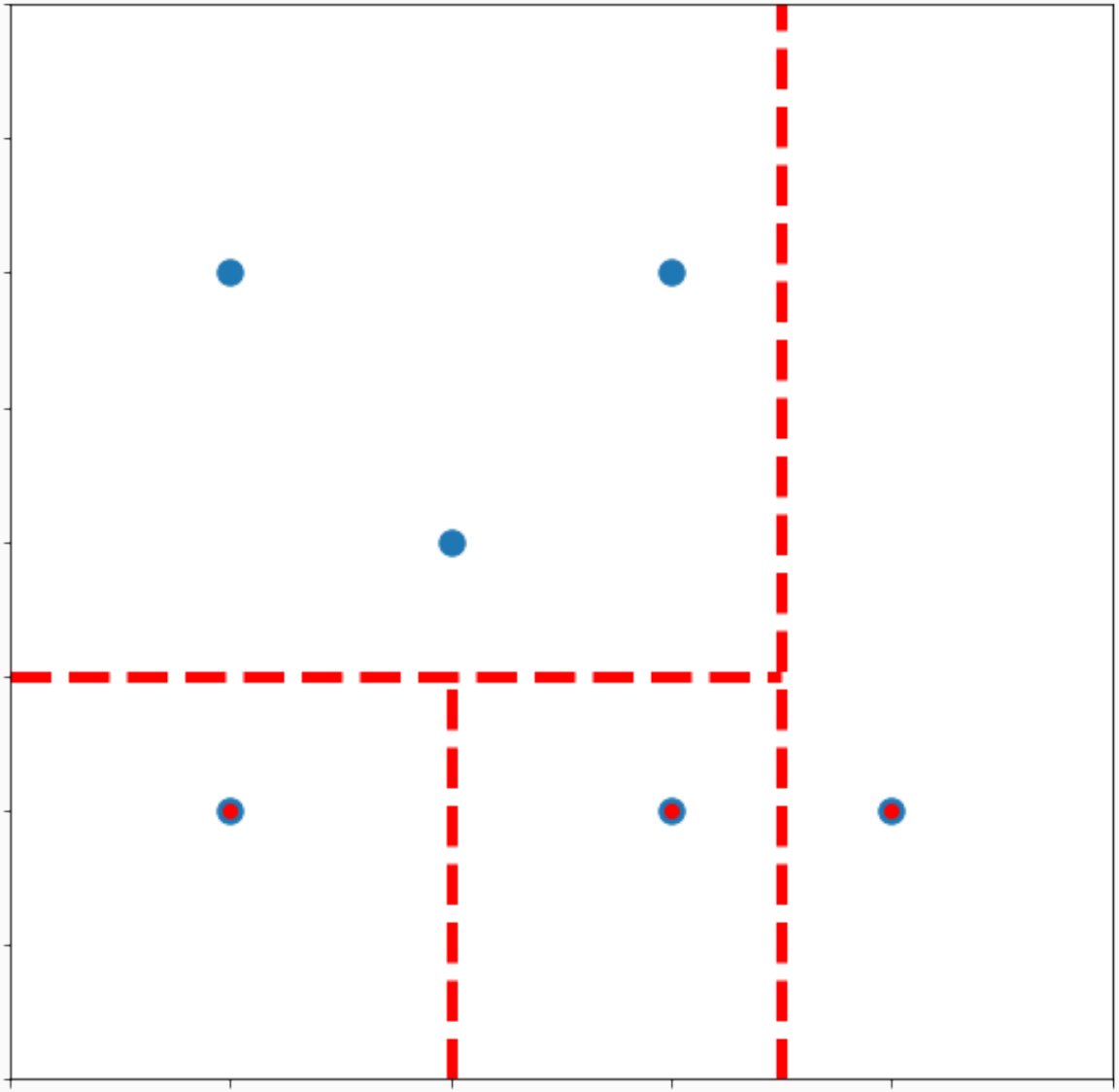
Isolation Forest



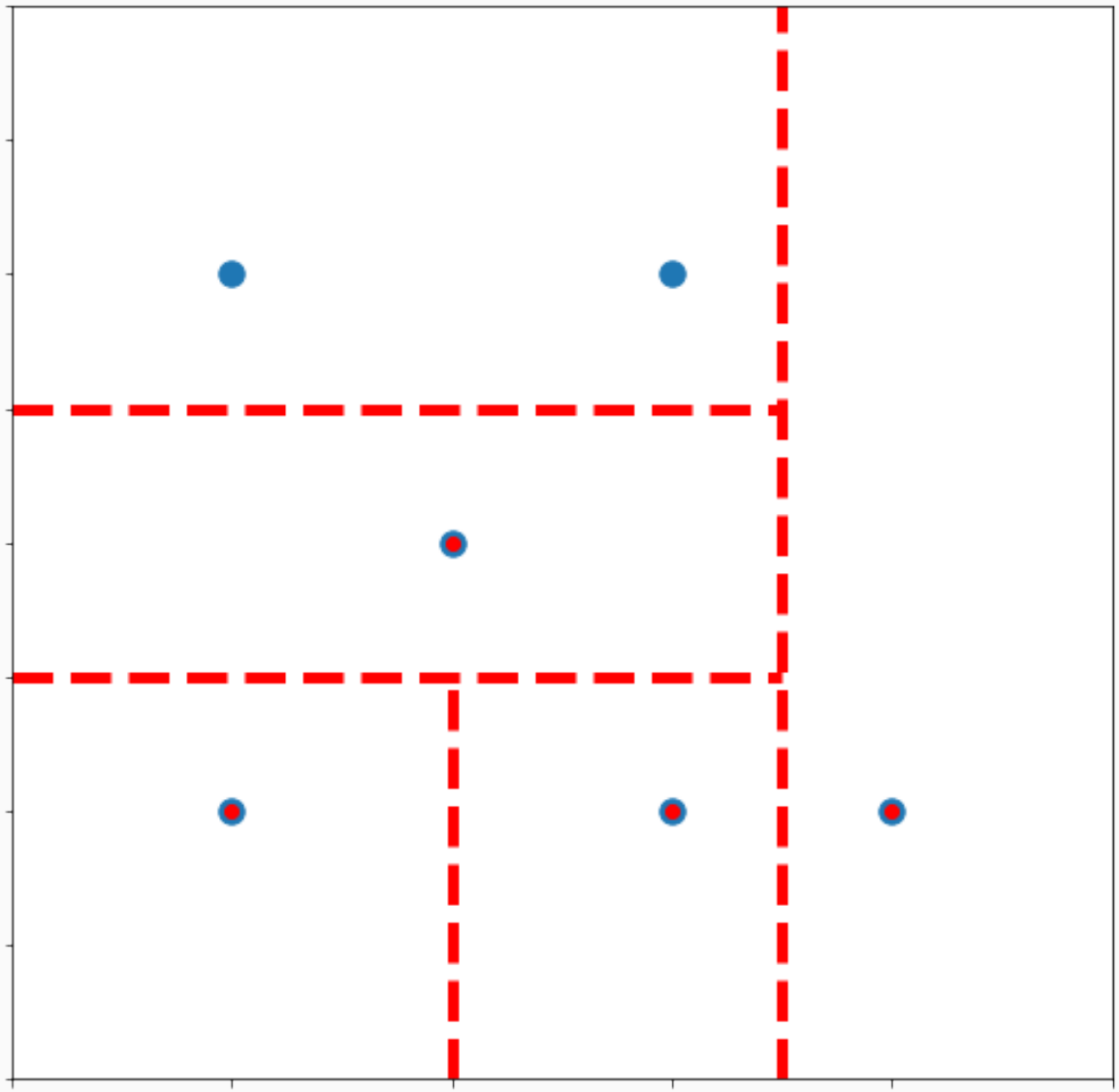
Isolation Forest



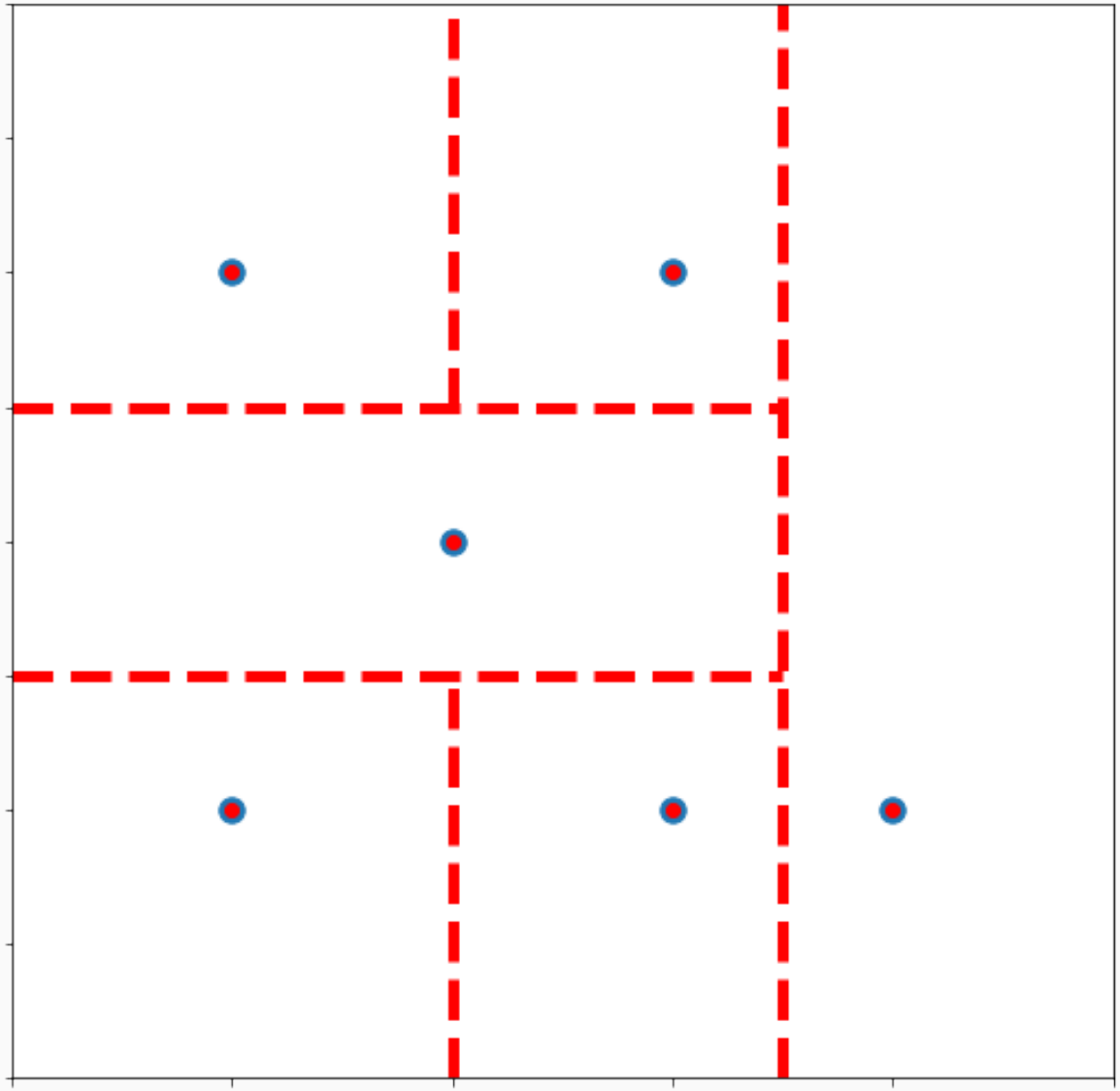
Isolation Forest



Isolation Forest

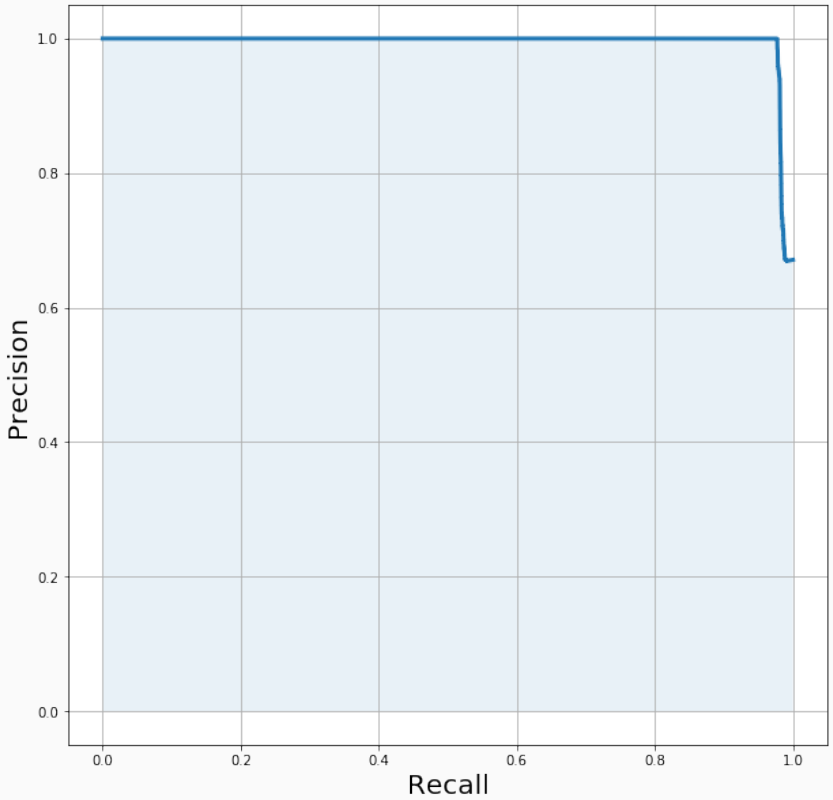


Isolation Forest



Isolation Forest

Время обучения	22s
Время на предсказание	28s
Precision/Recall AUC	0.994



Isolation Forest

Pros:

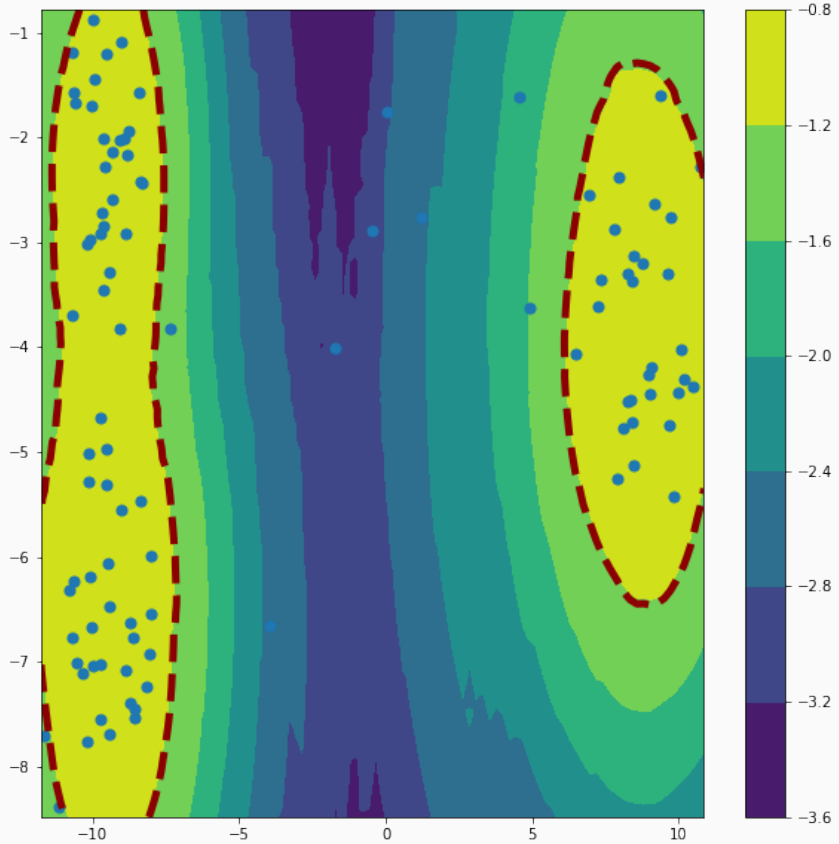
- Легко параллелится
- Робастные результаты

Cons:

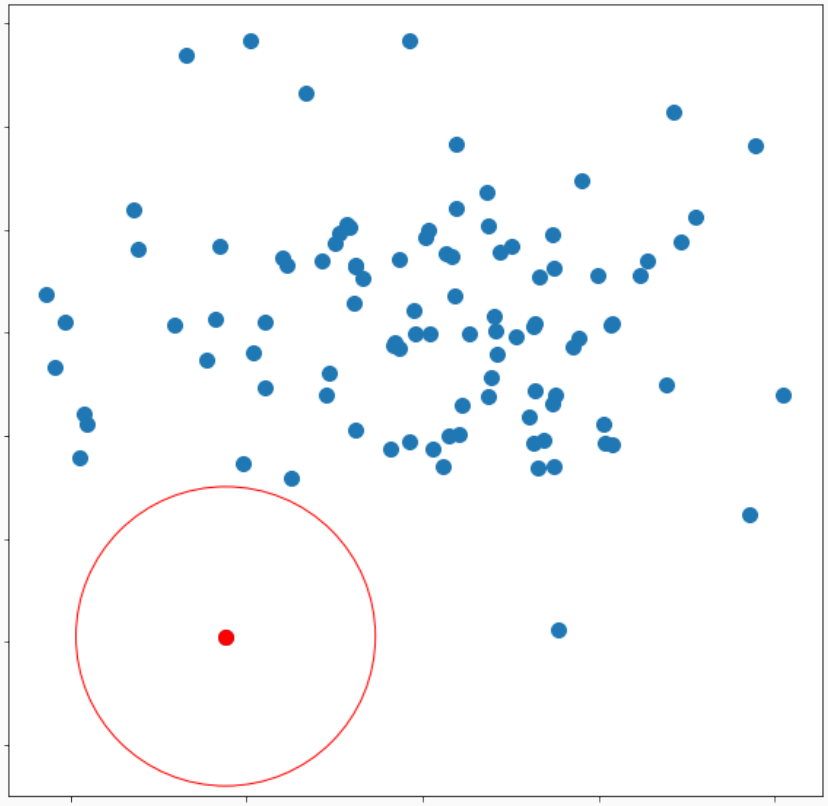
- Плохо интерпретируем

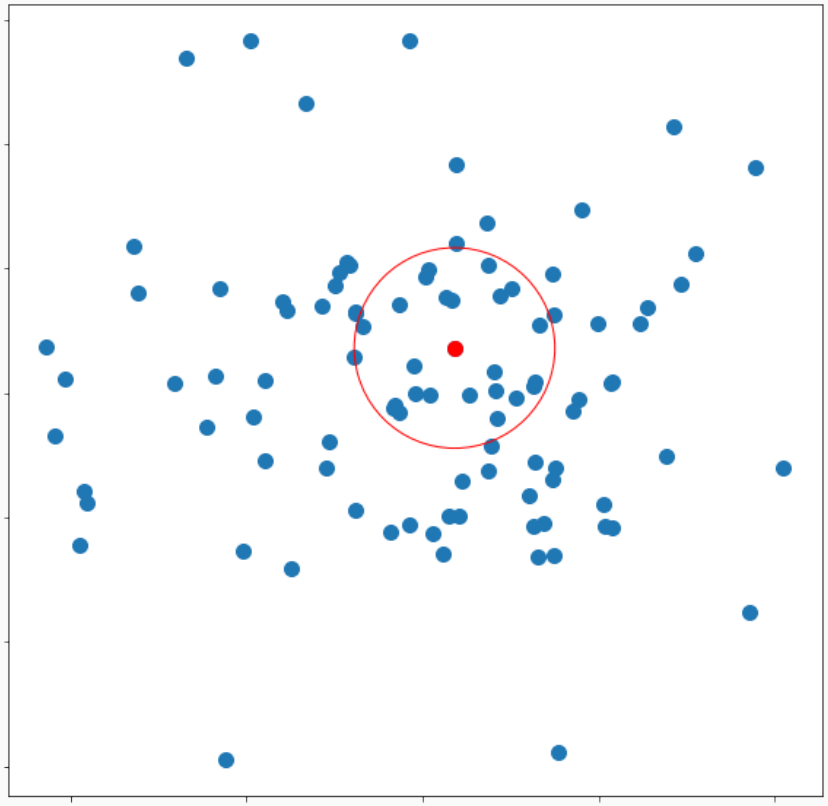
Local Outlier Factor

Оценивает локальную плотность на основе информации о ближайших соседях. Чем отдаленнее точка – тем более она аномальная



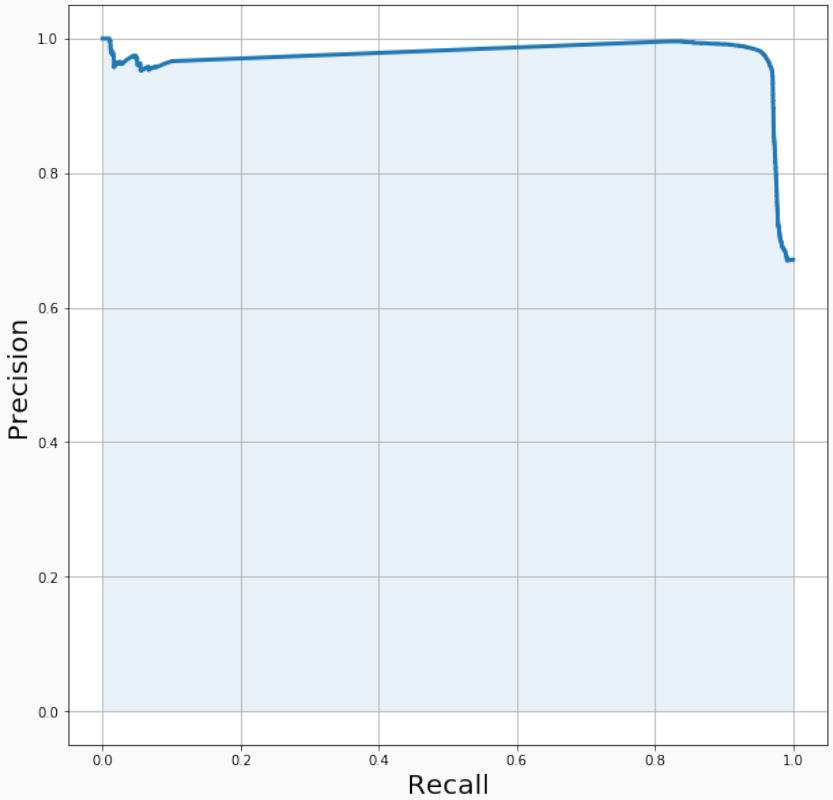
Local Outlier Factor





Local Outlier Factor

Время обучения	1s
Время на предсказание	40min
Precision/Recall AUC	0.984



Local Outlier Factor

Pros:

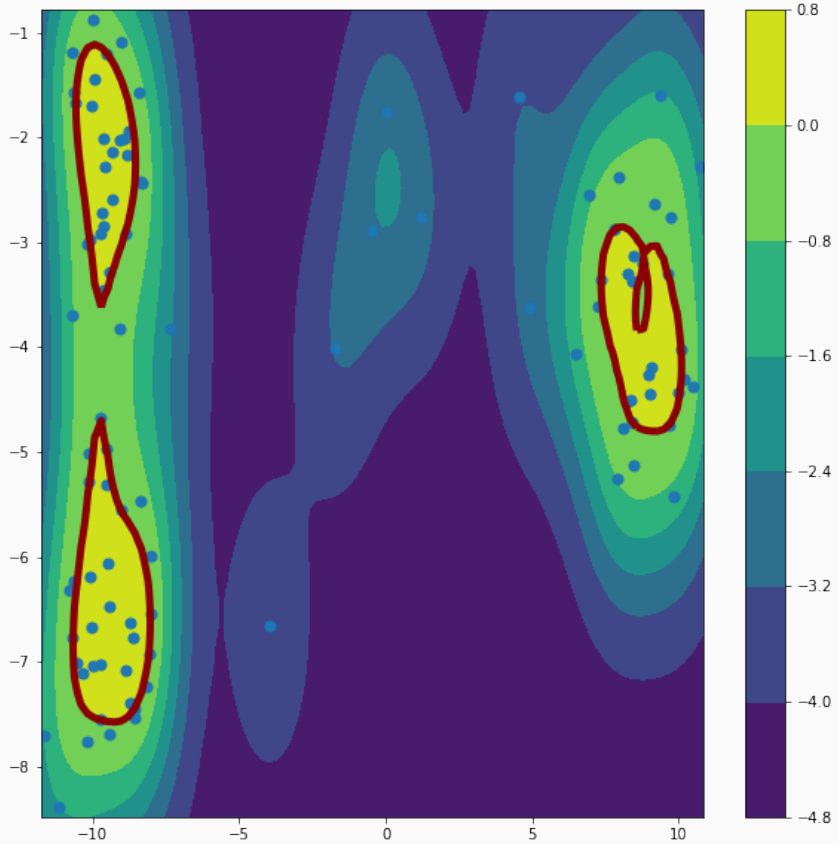
- Нет параметрических предположений
- Хорошо работает для низкоразмерных данных

Cons:

- Страдает от проклятья размерности
- Нужно хранить всю выборку
- Вычислительно сложный

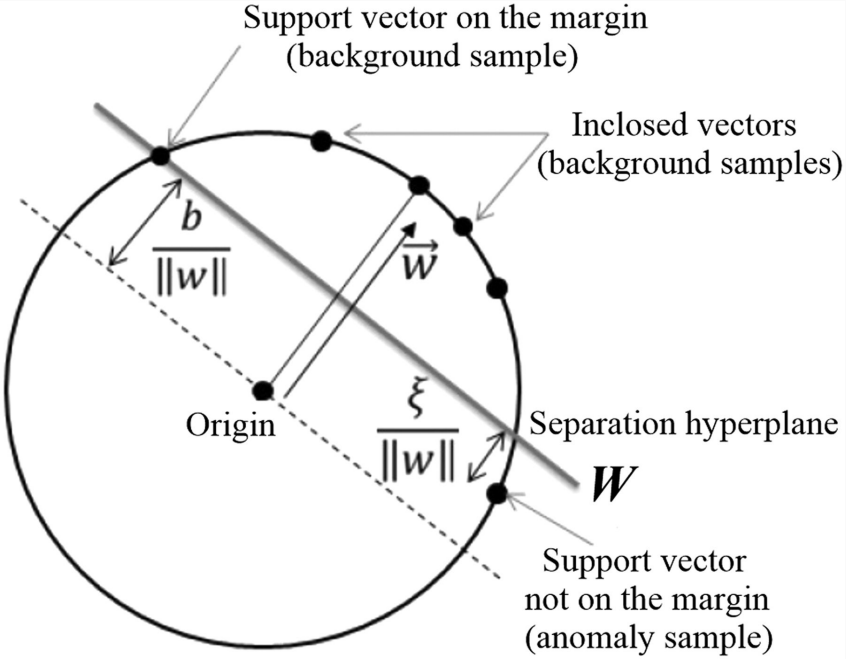
One Class SVM

Пытается отделить наблюдения от точки начала координат. Позволяет использовать ядровые методы



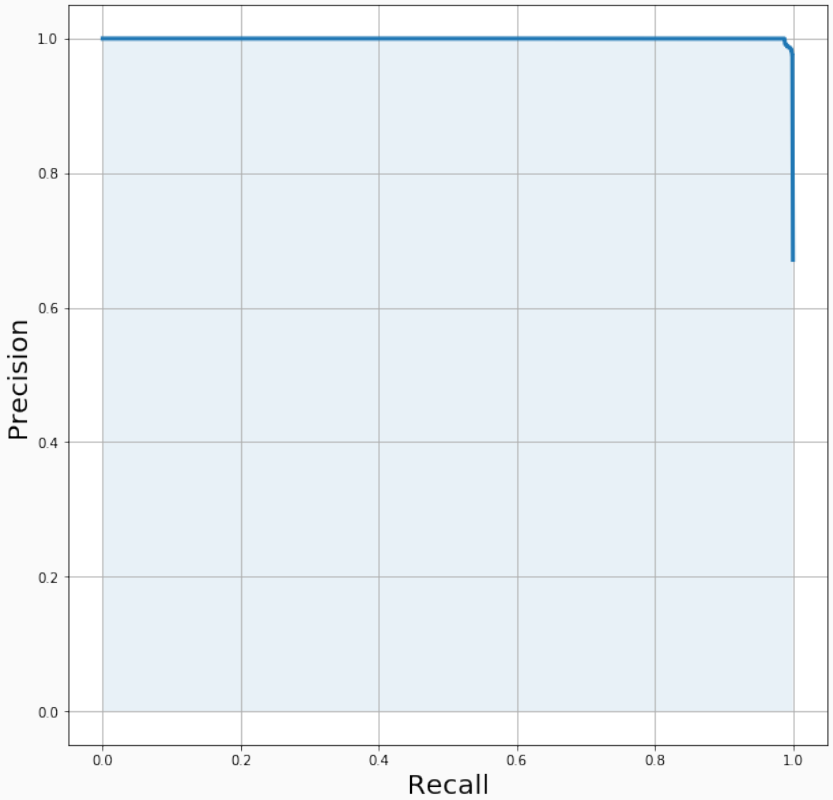
One Class SVM

Работает только в случае ядер, которые соответствуют отображению на поверхность некоторой гиперсферы



One Class SVM

Время обучения	1h 40m
Время на предсказание	1h
Precision/Recall AUC	0.999



One Class SVM

Pros:

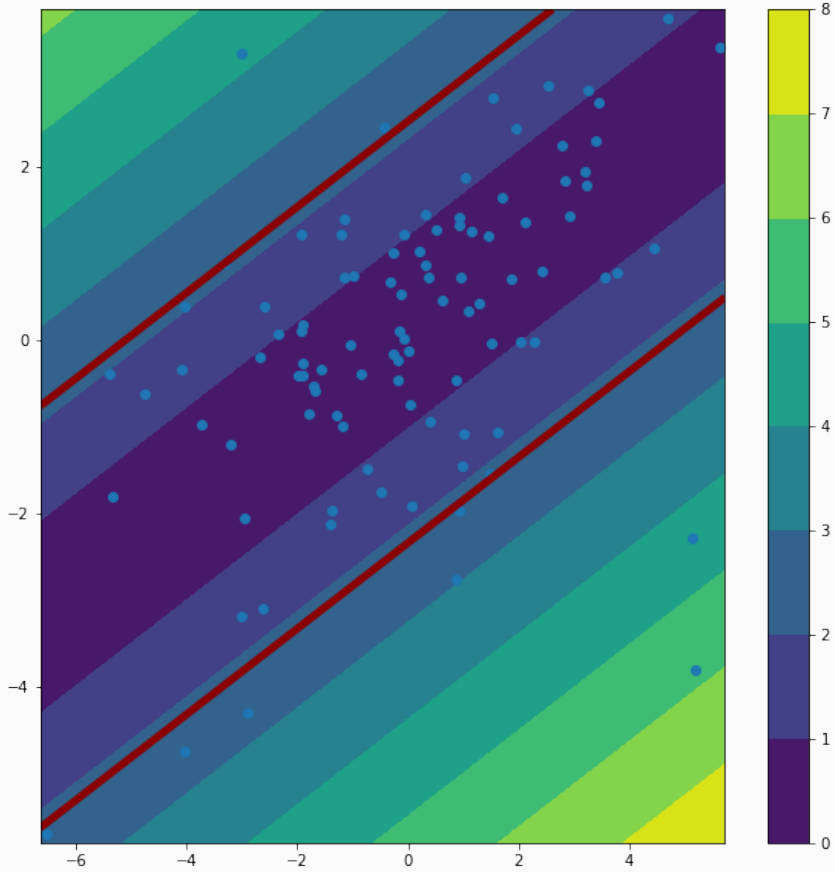
- Теоретическая обоснованность
- Нет явных предположений о распределении
- Можно применять не только к объектам из $\mathbb{R}^n$
- Результат сильно зависит от качества ядер

Cons:

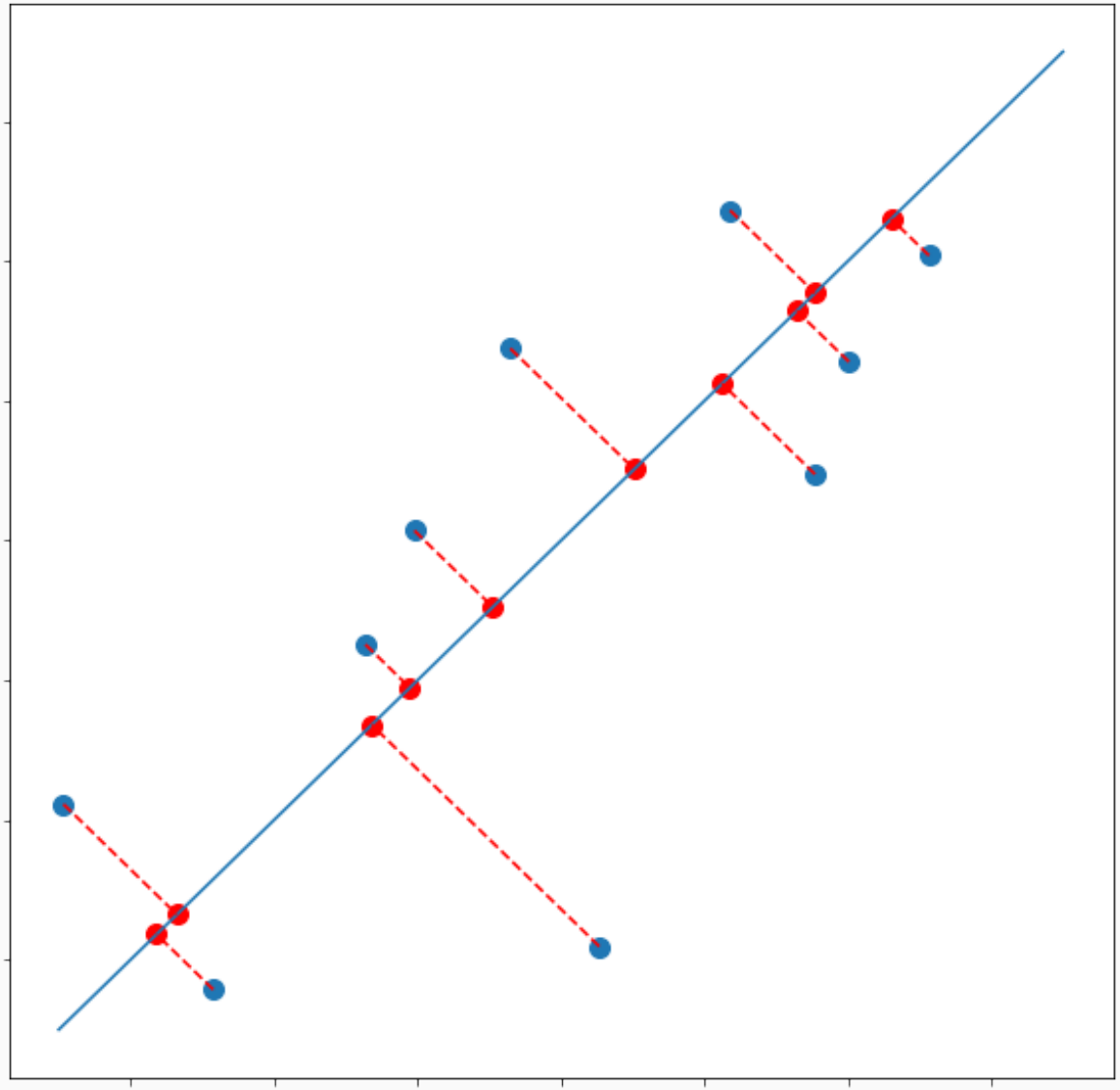
- Вычислительно сложный
- Приходится хранить часть выборки
- Результат сильно зависит от качества ядер

PCA

- Используем линейный метод снижения размерности
- Считаем расстояние от точек до линейного подпространства

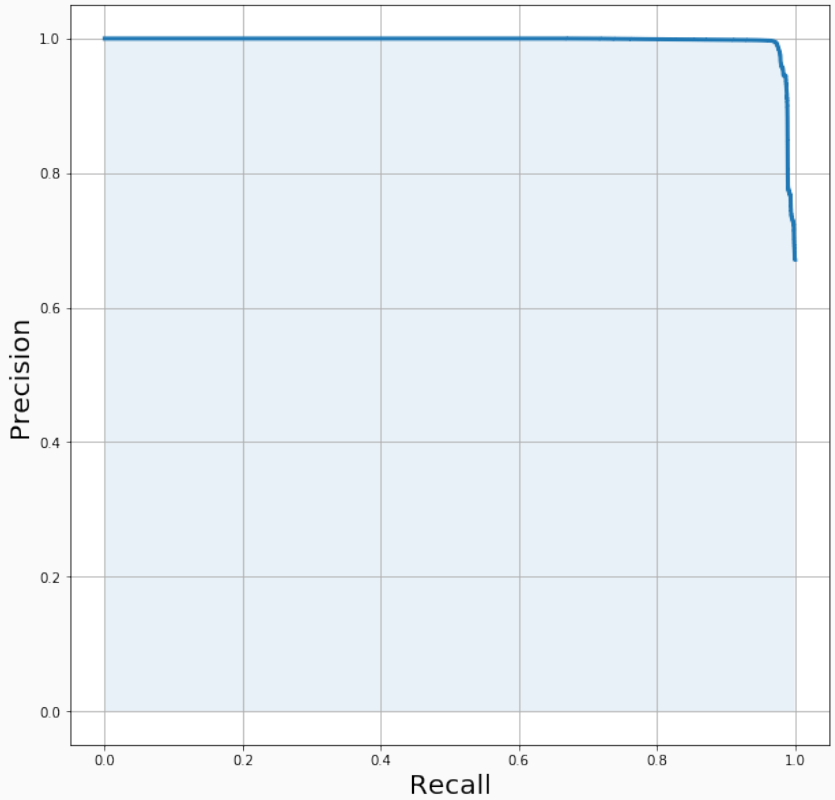


PCA



PCA

Время обучения	2.96c
Время на предсказание	2.02c
Precision/Recall AUC	0.996



Pro:

- Хорошо работает на высокоразмерных данных
- Эффективно считается
- Хорошо интерпретируется

Cons:

- Неявно подразумевает нормальное распределение
- Не работает для нелинейных случаев

SVDD+

Детектирования аномалий с применением привилегированной информации

- Обычные наблюдения $(x_1, \ldots, x_l)$
- Дополнительные наблюдения $(x_1^*, \ldots, x_l^*)$
- Обучение происходит на парах (x_i, x_i^*)
- Детектирование происходит только на (x_i)

KDD-99 Dataset

Можно выделить три группы признаков

TCP Dump	Экспертные признаки	Окно в 2 секунды
duration	login attempt	serror rate
protocol type	sudo attempt	same srv rate
flag	root login	diff srv rate
etc	etc	etc

SVDD+

Данные

- Базовые признаки использовались, как основное пространство
- Признаки на основе окна и советов экспертов использовались в качестве привилегированного пространства

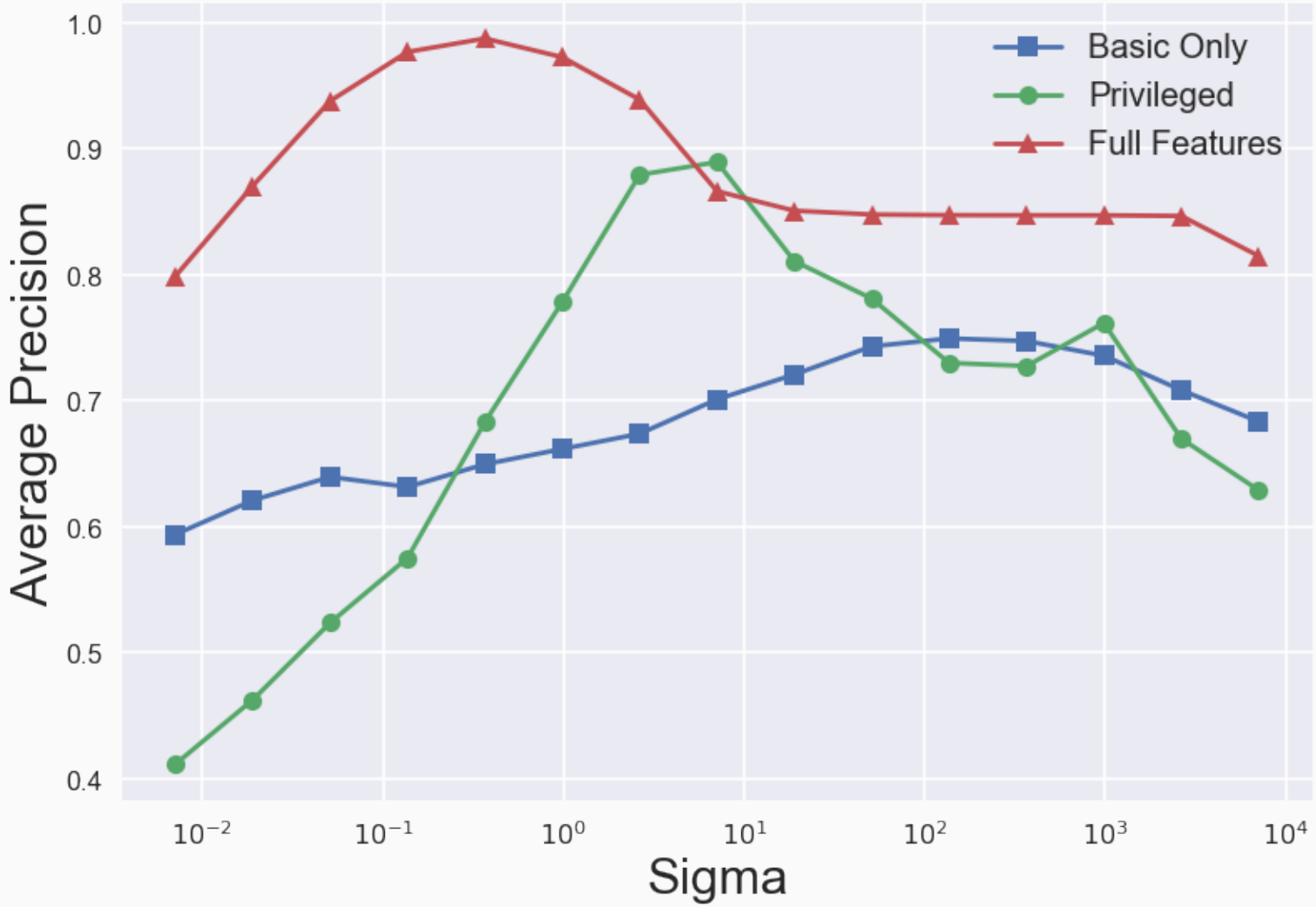
Параметры алгоритма

- Для всех экспериментов фиксируем $\nu = 0.1$
- Использовалась гауссово ядро

Оценка результатов

- Использовалась перекрестная проверка по пяти блокам
- Считаем AUC Precision/Recall

Результаты


Basic Only
Privileged
Full Features
Average Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
Sigma