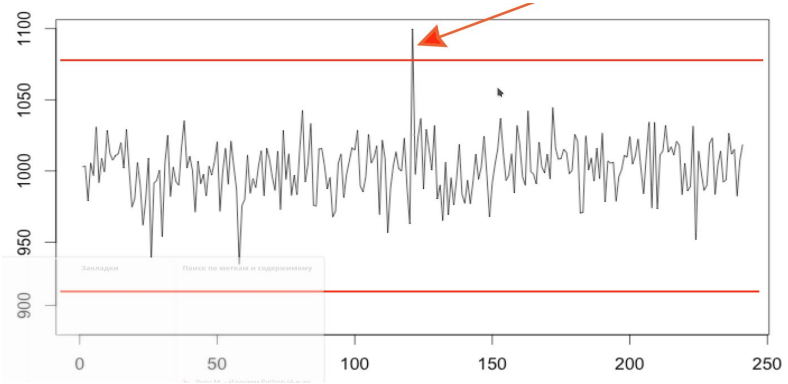
Об интеллектуальных методах обнаружения аномалий функционирования автоматизированных систем управления техническими процессами

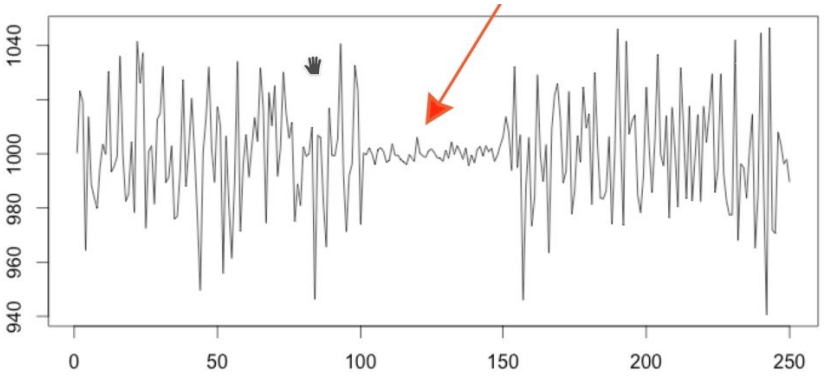
Что такое аномалии?

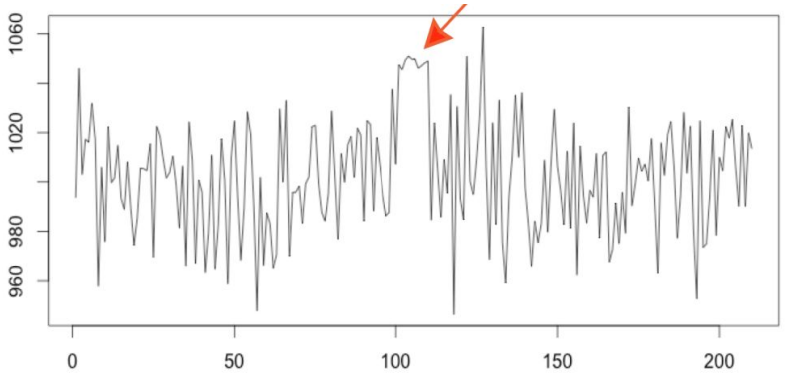
Перед началом, важно определиться с типами аномалий:
1. Точечные аномалии. Единичный случай аномального поведения, когда паттерн сильно отличается от всех предыдущих. Реальный пример: кража денег с карты, распознание на основе потраченных денег.
2. Контекстные аномалии. Аномалии завязанные на контекст, наиболее распространено в данных с временными рядами. Реальный пример: Траты $100 в день по праздникам нормально, но вызывает подозрения в других случаях.
3. Коллективные аномалии: множество данных, совокупно помогающих определить аномалии. Реальный пример: Попытка скопировать данные с удалённой машины.

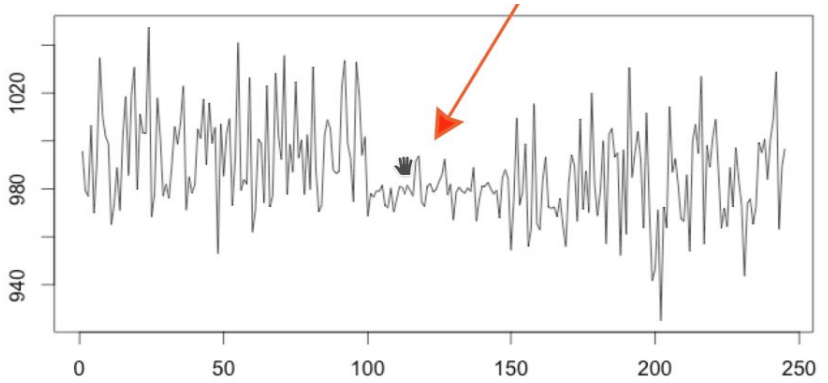
Простые статистические методы

- Простейшим способом обнаружения аномалий является выделение отклонений от обычных статистических параметров распределения, таких как среднее, медиана, мода и квантили.









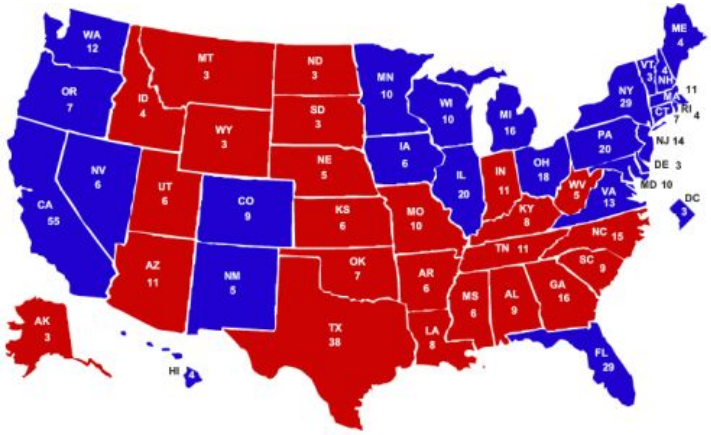
Минусы статистических подходов

Статистические методы плохо работают в следующих случаях:

- Данные содержат шум близкий к аномальному, бывает тяжело разделить нормальный и аномальный шум
- Плавное изменение данных может изменить данные распределения, поэтому такой способ не всегда применим
- Данные имеют сезонное распределение, это может потребовать разделение данных на несколько групп

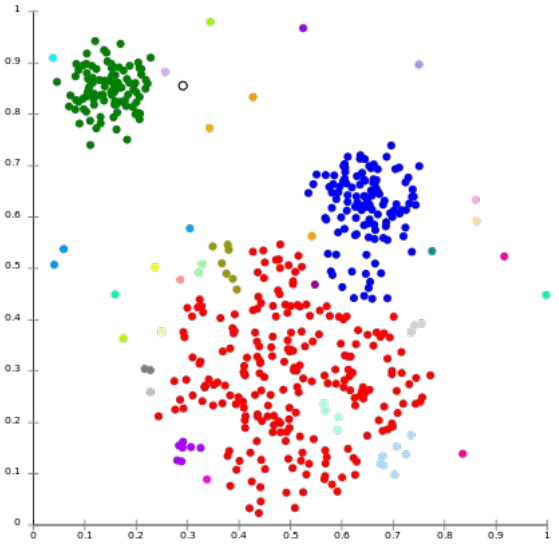
Поиск аномалий, основанный на плотности

Аномалии обнаруживаются из предположения, что схожие данные должны находится рядом. Метрики могут отличаться, в общему случае используется евклидова метрика, но в случае категориальных данных используются другие.



Кластеризация

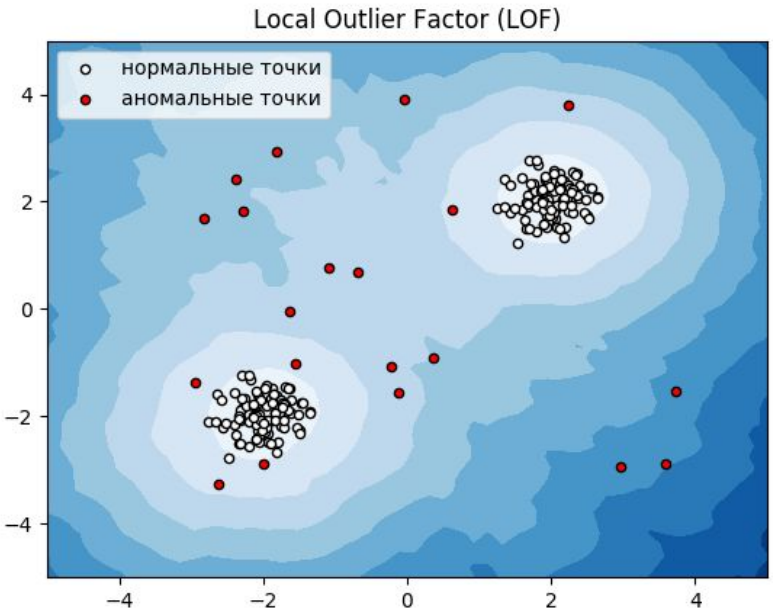
Ещё одной популярной техникой является кластеризация, основанная на обучении без учителя. Полагается, что схожие точки стремятся принадлежать к схожим кластерам.



- K — means — широко используемый алгоритм, он создаёт „k“ схожих кластеров данных. Точки, не относящиеся к кластерам

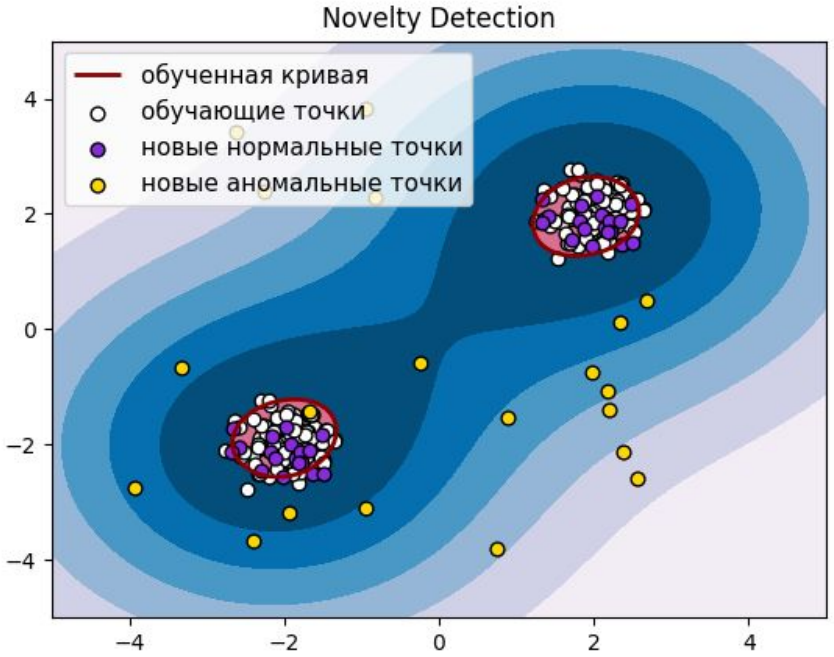
Local Outlier Factor

Фактор локального выброса основан на понятии локальной плотности, которая определяется k ближайшими соседями, расстояние до которых используется для оценки плотности. Сравнивая локальную плотность объекта с локальными плоскостями его соседей, можно идентифицировать точки, которые имеют



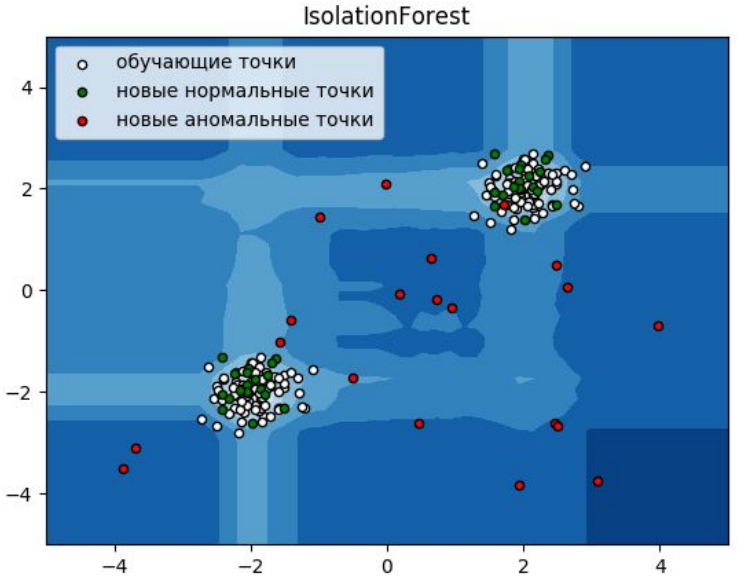
Метод опорных векторов(SVM)

- Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Обычно он используется при обучении с учителем, однако существуют



Изолирующий лес

- Ещё одним эффективным способом поиска аномалий является модифицированные алгоритм случайного леса Такой лес случайно выбирает признак, и затем выбирает случайное значение, по которому разделяет точки. Поскольку такое разделение можно представить в виде



Спасибо за внимание

Используемые статьи

http://scikit-learn.org/stable/modules/outlier_detection.html

https://www.datascience.com/blog/python-anomaly-detection

https://anomaly.io/anomaly-detection-normal-distribution/