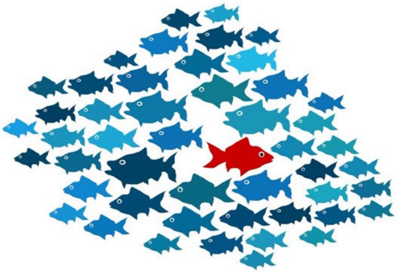
Обнаружение аномалий

Виктор Китов

victorkitov.github.io



Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Аномалии (выбросы)

- Аномалия (выброс, outlier) - объект, нетипичный для общего распределения объектов.
- Применения обнаружения аномалий (anomaly detection)
  - очистка данных (убрать ошибочные наблюдения)
  - обнаружение нетипичных объектов:
    - мошеннические транзакции в финансах
    - взлом компьютерной сети
    - мониторинг исправности устройств (станок, вертолет, ядерный реактор)
  - детектирование сдвига модели (concept drift)

Если есть разметка

- Если выбросы размечены в train, то это imbalanced class classification[1].
- Пусть $y = +1$ - редкий класс: $|n : y_n = +1| \ll |n : y_n = -1|$
- Как решать?

[1] Библиотека Python для несбалансированных классов.

Если есть разметка

- Если выбросы размечены в train, то это imbalanced class classification[1].
- Пусть $y = +1$ - редкий класс: $|n : y_n = +1| \ll |n : y_n = -1|$
- Как решать?
  - дублировать выбросы в выборке
  - обобщение: взвешенная ф-ция потерь $(w > 1)$

$$w \sum_{n:y_n=+1} \mathcal{L}\left(f_\theta\left(x_n\right), y_n\right) + \sum_{n:y_n=-1} \mathcal{L}\left(f_\theta\left(x_n\right), y_n\right) \to \min_\theta$$

  - исключить часть объектов класса -1
    - Алгоритм NearMiss: оставляем объекты -1 класса, ближайшие к объектам +1 класса
  - генерация синтетических объектов для выбросов
  - аугментация

[1] Библиотека Python для несбалансированных классов.

Генерация синтетических объектов[2]

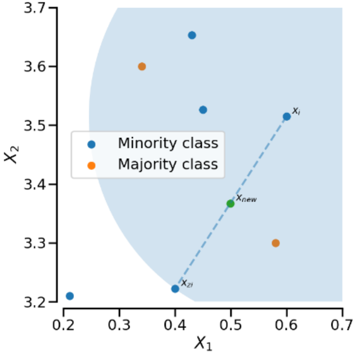
Метод SMOTE генерирует синтетич. объекты класса +1.

- для каждого объекта x_n с $y_n = +1$
  1. найдем K ближ. соседей $KNN(x_n)$
  2. P раз выберем случайные объекты из $KNN(x_n)$

$$A(x_n) = \{x_{i_1}, ... x_{i_K}\}$$

  3. для каждого $x' \in A(x_n)$ сгенерируем новый объект класса $y = +1$

$$x = (1-\alpha)\, x_n + \alpha x', \quad \alpha \sim U[0, 1]$$



[2]Как обобщить на категориальные признаки?

Расширение обучающей выборки

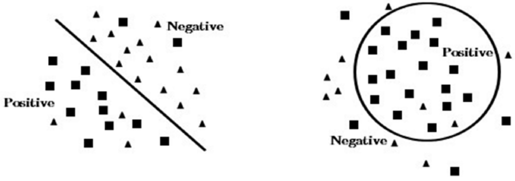
- Расширение обучающей выборки (data augmentation): модификации x, генерирующие реальные объекты того же класса.
- Как можно расширять выборку для
  - изображений
  - звуков
  - текстов

Расширение обучающей выборки

- Расширение обучающей выборки (data augmentation): модификации x, генерирующие реальные объекты того же класса.
- Как можно расширять выборку для
  - изображений
    - $\uparrow\downarrow$яркости, контраста
    - сдвиг / поворот с обрезкой
  - звуков
    - $\uparrow\downarrow$скорости
    - $+$помехи
    - $\uparrow\downarrow$ тембра (частоты)
  - текстов
    - замена слов синонимами
    - перевод на др. язык и обратно
    - суммаризация (прореживание предложений)

Методы обнаружения аномалий

- Обнаружение аномалий (anomaly detection) - обучение без учителя
  - нет разметки выбросов в train
  - но может быть в validation (для оценки)
- Несбалансированная классификация: есть примеры аномалий (есть паттерн)
  - выделяем область каждого класса
- Обнаружение аномалий: сложность в новизне (нет паттерна аномалии)
  - напр. детекция мошеннических действий (всё время новые)
  - выделяем область нормальности, остальное - выбросы



b. Classification   a. Anomaly detection

Методы обнаружения аномалий

- Методы оценивают степень нетипичности:

$$x \text{ - выброс} \Longleftrightarrow f(x) > threshold$$

- детекция выбросов (outlier detection): обучающая выборка содержит аномалии.
- детекция новизны (novelty detection): обучающая выборка не содержит аномалий.
  - выше пороги, чем в outlier detection
- Подходы:
  - статистический: $p(x) < t$
  - метрический: выброс далеко от др. точек
  - модельный: моделируем область нормальности

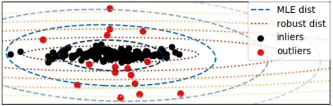
Содержание

Статистическое обнаружение аномалий

- Выбросы - точки с $p(x) < threshold$.
- Можем оценить $p(x)$ параметрически, например $\mathcal{N}(x|\mu,\Sigma) \propto e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$

1. Оценим $\widehat{\mu}, \widehat{\Sigma}$
2. $outlierness(x) = 1/p_{\widehat{\mu},\widehat{\Sigma}}(x)$

Mahalanobis distances of a contaminated data set:



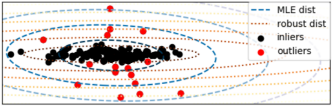
- В чем потенциальная проблема?

Статистическое обнаружение аномалий

- Выбросы - точки с $p(x) < threshold$.
- Можем оценить $p(x)$ параметрически, например $\mathcal{N}(x|\mu, \Sigma) \propto e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

1. Оценим $\widehat{\mu}, \widehat{\Sigma}$
2. $outlierness(x) = 1/p_{\widehat{\mu},\widehat{\Sigma}}(x)$

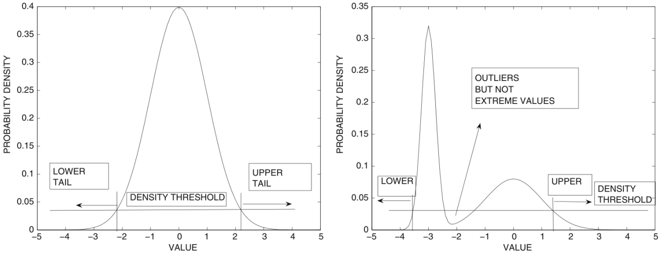
Mahalanobis distances of a contaminated data set:



- В чем потенциальная проблема?$\widehat{\mu}, \widehat{\Sigma}$ важно оценить устойчивым к выбросам способом.

Статистическое обнаружение аномалий

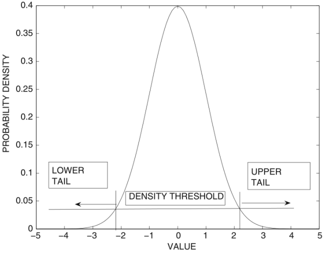
- Выбросы не обязательно на границе распределений:



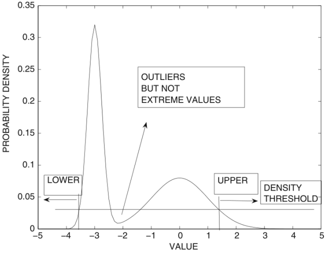
(a) Symmetric distribution (b) Asymmetric distribution

Статистическое обнаружение аномалий

- Выбросы не обязательно на границе распределений:



(a) Symmetric distribution



(b) Asymmetric distribution

- $p(x)$ можно оценить смесью распределений или KDE:

$$\widehat{p}(x) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{x - x_n}{h}\right)$$

Содержание

K-центров

- Если все точки train нормальные, то можем решить задачу K-покрытия:
    - найти $z_1, ...z_K$ такие, что

$$\min_k \rho\left(x_n, z_k\right) \le R \quad \forall n = 1, 2, ...N.$$

$$\text{outlierness}(x) = \min_k \rho\left(x, z_k\right)/R$$

- Это метод K-центров (K-centers[3]).
- Как связаны параметры K, R?

[3] Support Objects for Domain Approximation.

K-центров

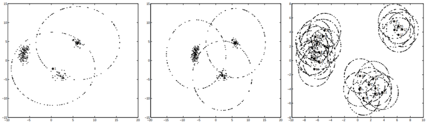
- Если все точки train нормальные, то можем решить задачу K-покрытия:
    - найти $z_1, ...z_K$ такие, что

$$\min_k \rho\left(x_n, z_k\right) \leq R \quad \forall n = 1, 2, ...N.$$

$$\text{outlierness}(x) = \min_k \rho\left(x, z_k\right)/R$$

- Это метод K-центров (K-centers[3]).
- Как связаны параметры K, R?

Связь K и R:



- Др. применение: опт. расположение K складов в N городах.

[3] Support Objects for Domain Approximation.

Алгоритм K-центров

```
выбираем z_1 случайным объектом
k:=1
ПОКА k ≤ K
    выбираем z_{k+1} самым удалённым объектом от {z_1, ...z_k}
    k:=k+1
```

Жадный алгоритм K-центров:

Перед наращиванием k можно пробовать улучшить расположение центров

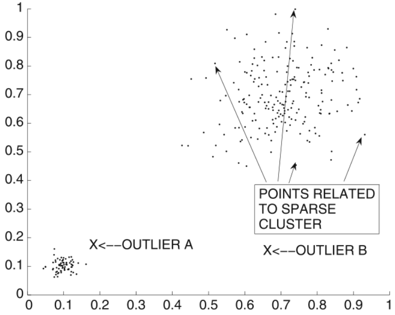
- z_k - наилучший объект в R-окрестности с точки зрения глобального R (сложность $O(N^2)$)

Но как быть, если обучающая выборка может содержать выбросы?

Обнаружение аномалий по расстоянию

Простые способы. Объект выброс, если расстояние выше порога

- $\rho(x, NN(x)) > t$ (до ближайшего соседа)
- $\min_k \rho(x, \mu_k) > t$ (до центра ближайшего кластера)



Но тогда выброс A либо пропущен, либо все точки разреженного кластера - выбросы.

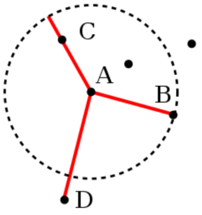
Метод local outlier factor

- Пусть $N_K(x)$ - множество K ближайших соседей x
- Определим ($NN_K(x)$ - K-й ближайший сосед x)

$$\rho_K(x) = \rho(x, NN_K(x)) \text{ (K-расстояние)}$$
$$\mathrm{rd}_K(x,z) = \max\{\rho(x,z), \rho_K(z)\}$$

- k-distance: устойчивее к случайным отдельным точкам.
- max: устойчивость к слишком близким точкам.

$$rd_K(A,B) = rd_K(A,C) < rd_K(A,D) \text{ для K=3}$$



Метод local outlier factor

- $\text{lrd}_K(x)$ (local reachability density)-плотность точек вокруг x:

$$\text{lrd}_K(x) = \frac{1}{\frac{1}{|N_k(x)|}\sum_{z\in N_k(x)} \text{rd}_K(x,z)}$$

- Метод local outlier factor - отношение плотности соседей x к плотности x:

$$LOF_K\left(x\right) = \frac{\frac{1}{|N_K(x)|}\sum_{z\in N_k(x)} lrd_K(z)}{lrd_K(x)}$$

- Это loss или score?

Метод local outlier factor

- $\text{lrd}_K(x)$ (local reachability density)-плотность точек вокруг x:

$$\text{lrd}_K(x) = \frac{1}{\frac{1}{|N_k(x)|}\sum_{z \in N_k(x)} \text{rd}_K(x,z)}$$

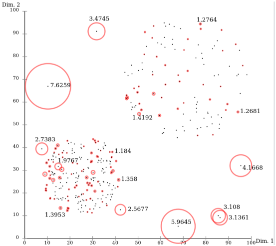
- Метод local outlier factor - отношение плотности соседей x к плотности x:

$$LOF_K(x) = \frac{\frac{1}{|N_K(x)|}\sum_{z \in N_k(x)} lrd_K(z)}{lrd_K(x)}$$

- Это loss или score?
- $LOF_K(x) \le 1$: типичная точка, $LOF_K(x) > 1$ - более удалённая.

Анализ

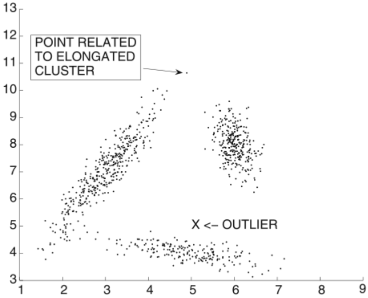
- LOF корректно выделяет выбросы как в контексте густых, так и в контексте разреженных соседей:



- Выброс, если $LOF_K(x) > t$: нужно подбирать t и K.
- Обобщается на др. $\rho(x, z)$
- Лучше работает с использованием метода случайных подпространств[4].

[4] Feature bagging for outlier detection.

Учет локального распределения точек



- Др. подход: учитывать локальное распределение точек.
- Подходы, учитывающие локальное распределение:
    - смесь Гауссиан
    - метод локального кластера (local cluster)
    - метод локальной окрестности (local neighborhood)

Метод локального кластера

1. Кластеризуем точки на K кластеров, используя расстояние Махаланобиса:
2. Для каждого кластера находим μ_k и Σ_k.
3. Для объекта x:
    1. находим ближайший кластер:

    $$\widehat{c} = \arg\min_c \sqrt{(x-\mu_c)^T \Sigma_c^{-1} (x-\mu_c)}$$

    2. степень нетипичности:

    $$\text{outlierness}(x) = \sqrt{(x-\mu_{\widehat{c}})^T \Sigma_{\widehat{c}}^{-1} (x-\mu_{\widehat{c}})}$$

Метод локальной окрестности

1. Инициализируем $L_K(x) = \{x\}$
2. Для $k = 1, 2, ...K$:
    1. $x_k = \arg\min_z \rho(z, L_K(x))$
    2. $L_K(x) := L_K(x) \cup \{x_k\}$
3. Исключим x: $L_K(x) := L_K(x) \backslash \{x\}$
4. Используя $L_K(x)$ рассчитаем $\mu(x)$ и $\Sigma(x)$
5. Степень нетипичности:

$$\text{outlierness}(x) = \sqrt{(x - \mu(x))^T \Sigma(x)^{-1} (x - \mu(x))}$$

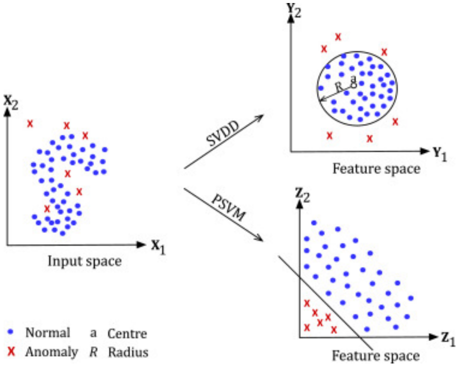
Вычислительно сложнее, зато лучше учитывает распределение вокруг x.

Содержание

3 Модельные методы
- Одноклассовый метод опорных векторов
- Изолирующий лес

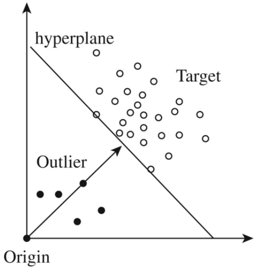
Одноклассовый метод опорных векторов[5]

- Преобразование пр-ков: $x \to \Phi(x) \in F$.
- Отделим в пространстве $\Phi(x)$ нормальные точки от остальных
  - гиперплоскостью, максимально отдалённой от нуля
  - шаром минимального радиуса



[5] Estimating the Support of a High-Dimensional Distribution.

Отделение гиперплоскостью



$$\begin{cases} \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_{n=1}^{N}\xi_n - \rho \to \min_{w,\xi\in\mathbb{R}^N,\rho\in\mathbb{R}} \\ \langle w, \Phi(x_n)\rangle \geq \rho - \xi_n; \quad \xi_n \geq 0, \quad n = 1, 2, \ldots N. \end{cases}$$

$$f(x) = \text{sign}\left(\langle w, \Phi(x)\rangle - \rho\right) \qquad \text{-1 для выброса}$$

Максимизируем расстояние от нуля до гиперплоскости $\frac{\rho}{\|w\|}$.

Гиперпараметр $\nu \in (0, 1)$ - макс. доля выбросов в выборке.

Отделение гиперплоскостью - решение

В терминах ядер $K(x,z)=\langle\Phi(x),\Phi(z)\rangle$:

$$f(x)=\text{sign}\left(\sum_n \alpha_n K(x_n,x)-\rho\right)$$

где $\{\alpha_n\}$ находятся из решения двойственной задачи:

$$\begin{cases}\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j)\to\min_\alpha\\ 0\le\alpha_i\le\frac{1}{\nu N};\quad \sum_i\alpha_i=1\end{cases}$$

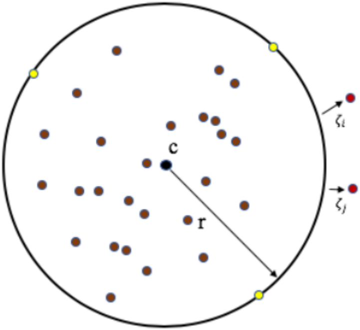
- $\alpha_i=0$: обычные объекты
- $\alpha_i\in\left(0,\frac{1}{\nu N}\right)$: на гиперплоскости, по 1 такому x_i находим:

$$\rho=\langle w,\Phi(x_i)\rangle=\sum_n\alpha_n K(x_n,x_i)$$

- $\alpha_i=\frac{1}{\nu N}$: выбросы[6]

[6]Как отсюда следует, что ν-макс. доля выбросов в выборке?

Отделение шаром



$$\begin{cases} R^2 + \frac{1}{\nu N}\sum_n \xi_n \to \min_{R\in\mathbb{R},\xi\in\mathbb{R}^N,c\in F} \\ \left\|\Phi\left(x_n\right) - c\right\|^2 \le R^2 + \xi_n; \quad \xi_n \ge 0; \quad n = 1, 2, \dots N. \end{cases}$$

$$f(x) = \text{sign}\left(R^2 - \left\|\Phi\left(x_n\right) - c\right\|^2\right)$$ -1 для выброса

Отделение шаром - решение

В терминах ядер $K(x,z)=\langle\Phi(x),\Phi(z)\rangle$:

$$f(x)=\text{sign}\left(R^2-\sum_{i,j}\alpha_i\alpha_j L(x_i,x_j)+2\sum_i\alpha_i K(x_i,x)-K(x,x)\right)$$

где $\{\alpha_n\}$ находятся из решения двойственной задачи:

$$\begin{cases}\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j)-\sum_i\alpha_i K(x_i,x_i)\to\min_\alpha\\ 0\le\alpha_i\le\frac{1}{\nu N};\quad\sum_i\alpha_i=1\end{cases}$$

- $\alpha_i=0$: обычные объекты
- $\alpha_i\in\left(0,\frac{1}{\nu N}\right)$: на гиперплоскости, по 1 такому x_i находим:

$$R^2=\|\Phi(x_i)-c\|^2=K(x_i,x_i)-2\sum_n\alpha_n K(x_n,x_i)+\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j)$$

- $\alpha_i=\frac{1}{\nu N}$: выбросы[7]

[7] Как отсюда следует, что ν-макс. доля выбросов в выборке?

Ядерное обобщение с RBF ядром

популярные ядра: $K(x,z)=e^{-\frac{\|x-z\|^2}{2\sigma^2}}, \quad K(x,z)=(\langle x,z\rangle+1)^K$

- Стационарные ядра[8]: $K(x,z)=G(x-z)$
  - инвариантны к сдвигу $K(x,z)=K(x+\Delta,z+\Delta)$
  - переводят $\{x_n\}$ на сферу

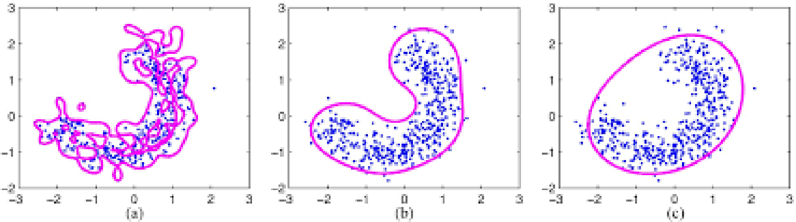
$$\|\Phi(x_n)\|^2=\langle\Phi(x_n),\Phi(x_n)\rangle=K(x_n,x_n)=G(0)=\text{const}$$

  - для данных на сфере минимизация сегмента (отделение гиперплоскостью) и минимизация шаром эквивалентны => 2 последних метода дают одинаковый результат.

[8] Будут ли Гауссово и полиномиальное ядра стационарными?

Параметры

Одноклассовый SVM с RBF ядром ($\sigma\uparrow$)



ν контролирует размер фигуры (и долю выбросов).

3 Модельные методы
- Одноклассовый метод опорных векторов
- Изолирующий лес

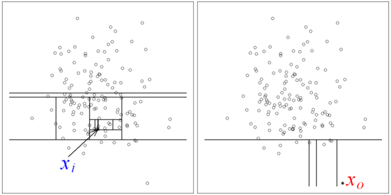
Изолирующее дерево

```
инициализировать корень всеми наблюдениями

ПОКА (существуют узлы с несовпадающими наблюдениями
      глубины < S):   # рекомендуется S=8
   выбрать такой узел
   выбрать случайный неконстантный признак f ∈ [f_min, f_max]
   выбрать случайный порог t ∈ (f_min, f_max)
   разбить узел на 2 подузла по правилу f ≤ t
```

Построение изолирующего дерева (isolation tree).

- Дерево строится без учителя.
- Как по нему оценить типичность x?



(a) Isolating x_i (b) Isolating x_o

Изолирующий лес

- Типичность объекта в дереве

$$h(x) = p(x) + c(m)$$

- $p(x)$ - глубина пути в дереве
    - m=#др. объектов в листе
    - $c(m) = 2\left(\ln\left(m-1\right) + 0.57\right) - 2\left(m-1\right)/m$ - оценка доп. пути до x, если бы дерево строилось до конца.
- По одному дереву считать нельзя (много случайности).

Изолирующий лес

- Типичность объекта в дереве

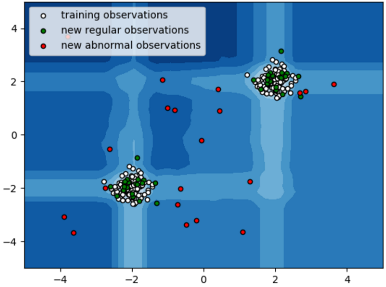
$$h(x) = p(x) + c(m)$$

- $p(x)$ - глубина пути в дереве
    - m=#др. объектов в листе
    - $c(m) = 2(\ln(m-1) + 0.57) - 2(m-1)/m$ - оценка доп. пути до x, если бы дерево строилось до конца.
- По одному дереву считать нельзя (много случайности).
- Изолирующий лес (isolation forest) - ансамбль K независимых изолирующих деревьев (рекоменд. $K = 100$).

$$\text{outlierness}(x) = 2^{-\frac{\mathbb{E}\{h(x)\}}{c(N)}}, \quad N = \#\text{объектов обуч. выборки}$$

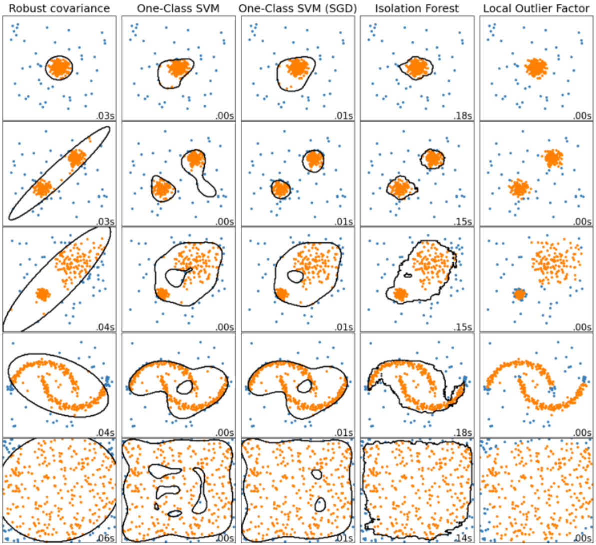
- outlierness$(x) \approx 1$: выброс
- outlierness$(x) \approx 0.5$: обычный объект

Преимущества



- ⊕ : Работает с вещественными, порядковыми, бинарными признаками.
- ⊕ : Быстрый, интерпретируемый алгоритм.
- ⊕ : Интерпретируемая outlierness$(x) \in (0, 1)$
- ⊕ : Обучается, даже если в X нет выбросов.
    - в отличие от одноклассового SVM
- ⊕ : Хорошо учится на малой подвыборке типичных объектов.

Сравнение методов

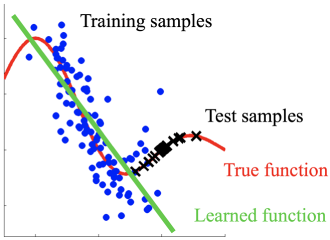


Содержание

Сдвиг модели

Связанно с обнаружением аномалий - сдвиг модели (concept drift): изменяется целевая зависимость $y = f(x)$

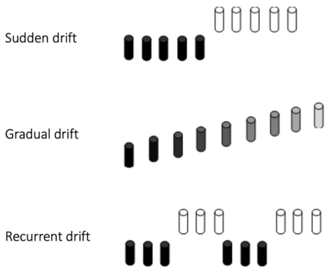
- изменения носят устойчивый характер во времени
  - у клиента изменились вкусы, женился, переехал
  - изменилась внешняя ситуация (пандемия)



- Будем наблюдать устойчивые изменения в ошибках модели.

Сдвиг модели

- Каждый тип сдвига детектируется и обрабатывается по-своему:



- В отличие от аномалий, сдвиги носят устойчивых характер.
- Детекция - на основе статистик сравнения распределений ошибок во времени.

Заключение

- Детекция выбросов - задача обучения без учителя
  - если с учителем - то это классификация несбалансированных классов
    - генерация объектов для редкого класса
    - удаление объектов для частого класса
- Оценка - по размеченной валидации, используя ROC, AUC.
- Методы обнаружения аномалий:
  - статистические: $p(x) < t$
  - метрические: выброс далеко от др. точек
  - модельные: моделируем область нормальности
    - one-class SVM
    - isolation forest