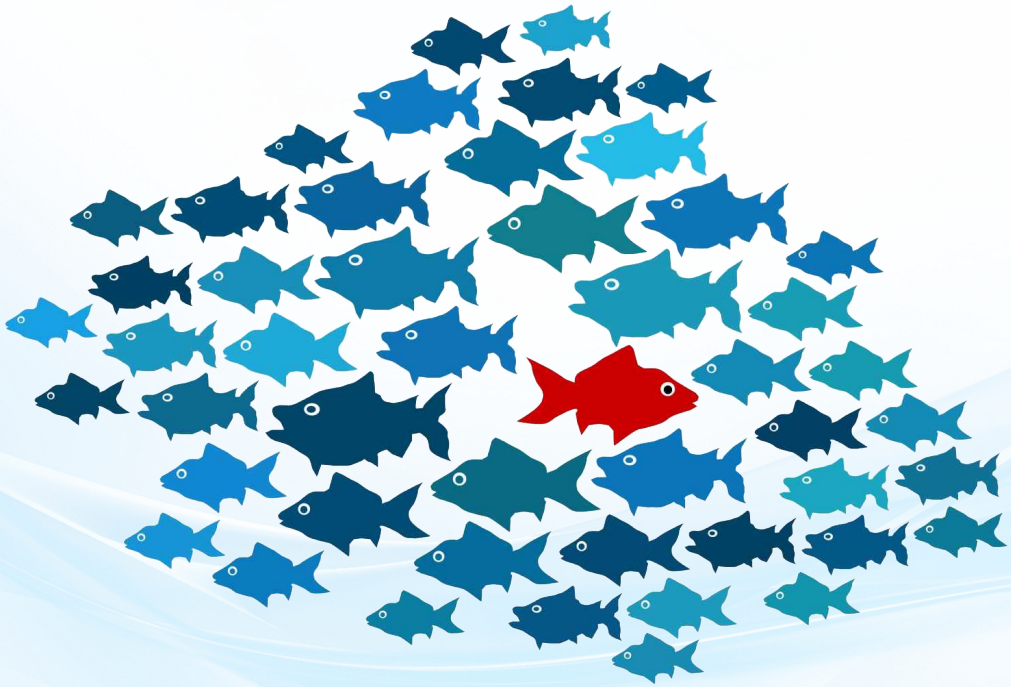
Доп. задачи анализа данных

8 мая 2020

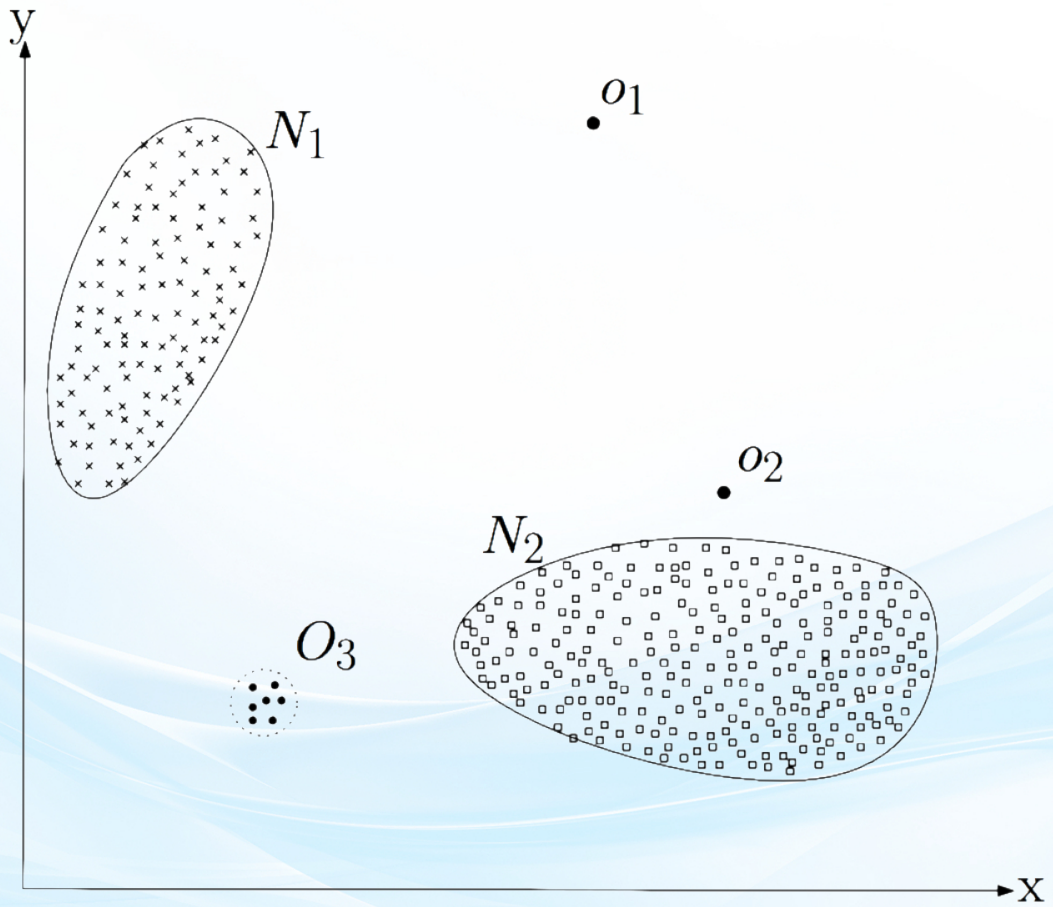
Аномалии



Аномалии

Аномалия — паттерн в данных,

который отличается от обычного поведения.



Аномалии (Outlier)

Аномалии разделяют на 2 группы — выбросы и новизну.

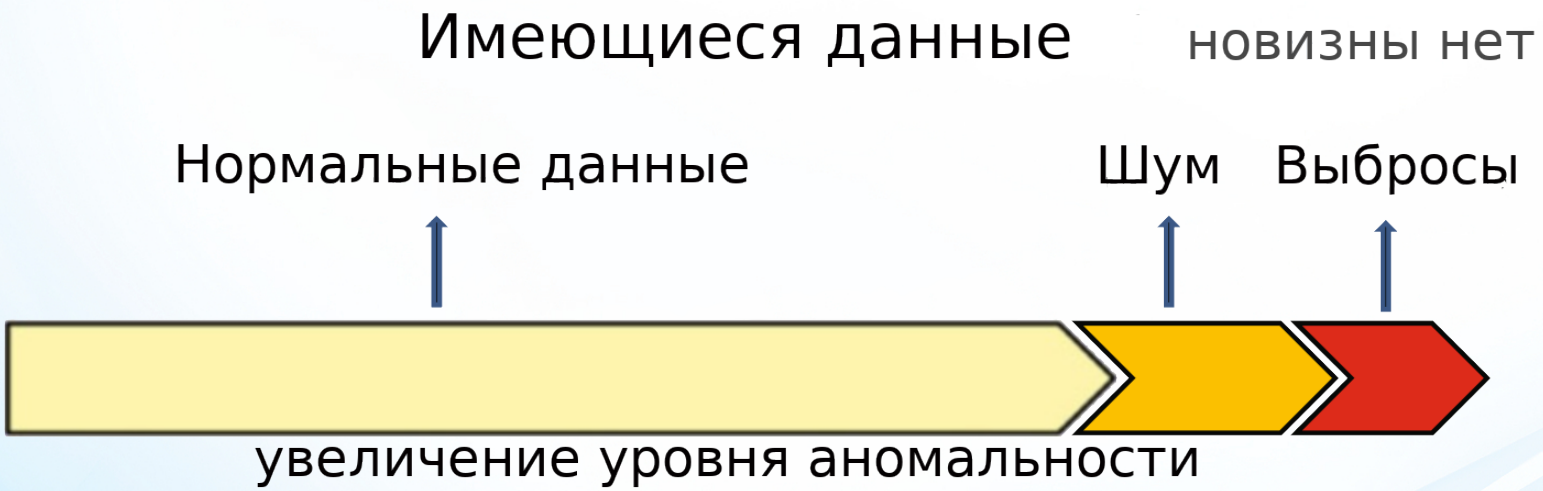
Выбросы — аномалии, присутствующие в обучающих данных.

Выбросы могут появиться на тестируемых данных.

Новизна — аномалии, которых нет в обучающих данных.

Новизна может появиться только на тестируемых данных.

Аномалии



Новизна может встретиться только в новых данных.

Аномалии.Выбросы

Причины появления выбросов

Ошибки в данных:

1. неточные измерения, ошибки записи при записи, округление;
2. показания сломанных приборов;
3. ошибки других моделей.

Основное применение методов обнаружения выбросов

Чистка данных для дальнейшего анализа модели.

Аномалии.Новизна

Причины появления новизны

Совершенно новые воздействия на систему, например:

1. вирус атаковал операцинную систему;
2. вор воспользовался чужыми банковскими картами и потратил все деньги на них;
3. сломался прибор и стал давать неправильные показания;
4. пациент заболел и у него изменился сердечный ритм;
5. начал пробуждаться давно потухший вулкан.

Аномалии.Новизна

Применение

1. Распознавание вредоносной активности в компьютерных системах;
2. Обнаружение подозрительных банковских операций;
3. Детектирование поломок приборов;
4. Медицинская диагностика;
5. Сейсмология.

Актуальность

Поиск в Google Scholar по статьям начиная с **2010** года:

- anomaly detection — **366 000** статей;
- outlier detection — **146 000** статей;
- novelty detection — **122 000** статей;
- обнаружение аномалий — **16 200** статей.

Поиск в Google Scholar по статьям начиная с **2015** года:

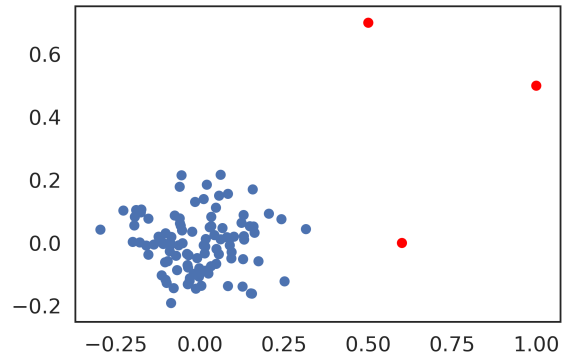
- anomaly detection — **128 000** статей;
- outlier detection — **51 600** статей;
- novelty detection — **34 400** статей;
- обнаружение аномалий — **14 100** статей.

Типы аномалий

Индивидуальная

точка аномальная по отношению

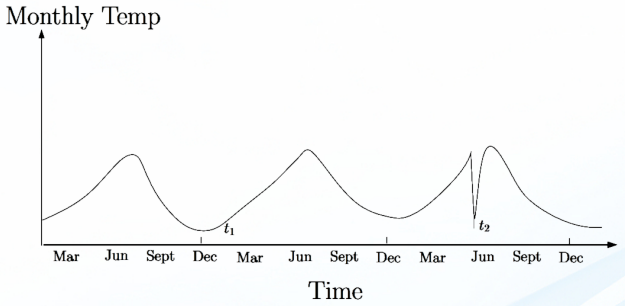
ко всем остальным данным



Контекстная

точка аномальная по отношению

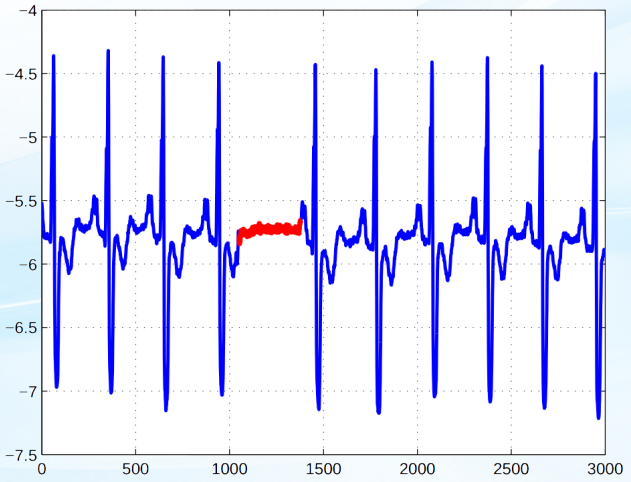
к своему контексту



Коллективная

группа точек аномальна

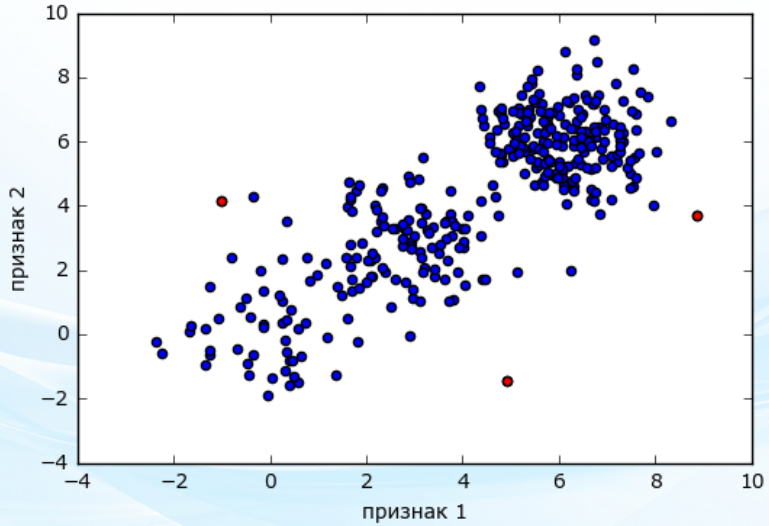
по отношению к остальным данным



Детектирование аномалий

Сложности детектирования аномалий

- То, какой объект считать аномальным, а какой нет, сильно зависит от задачи.
- Часто нет размеченных данных, тяжело определить качество модели.
- Аномалии могут не выделяться по отдельным признакам.



Типы задач

1. *Обучение с учителем*:
   Выборка полностью размечена на типичные и аномальные.
   Особенность — сильный дисбаланс классов.
2. *Частичное обучение с учителем*:
   Выборка частично размечена.
   Например, известно, что в некоторые моменты прибор ломался.
3. *Обучение без учителя*:
   Выборка не размечена, но предполагается,
   что типичных объектов существенно больше аномальных.

На выходе

1. Скоры — оценка степени аномальности объекта;
2. Метки — аномальный или нет.

Подходы к детектированию аномалий

1. Классификация — мультиклассовая, одноклассовая (One Class SVM), нейронные сети;
2. Основанные на ближайших соседях — расстояние до k-го соседа, LocalOutlierFactor;
3. Основанные на кластеризации — DBSCAN, ROCK, SNN. Двух этапные — SOM, k-means, EM. Cluster-Based Local Outlier Factor;
4. Статистические методы — ящик с усами, Grubb's test, регрессия, гистограммы, KDE;
5. Спектральный анализ — PCA;
6. Случайные леса — iForest (IF), RRCF;
7. Контекстная аномальность — ARIMA, HMM;
8. Коллективная аномальность — HMM, WCAD, MEMM, IMM;
9. Нейронные сети — автоэнкодеры, GAN, байес. нейронные сети.

Что изучим сегодня

Для всех методов:

- обучение без учителя;
- индивидуальные аномалии.

Подходы:

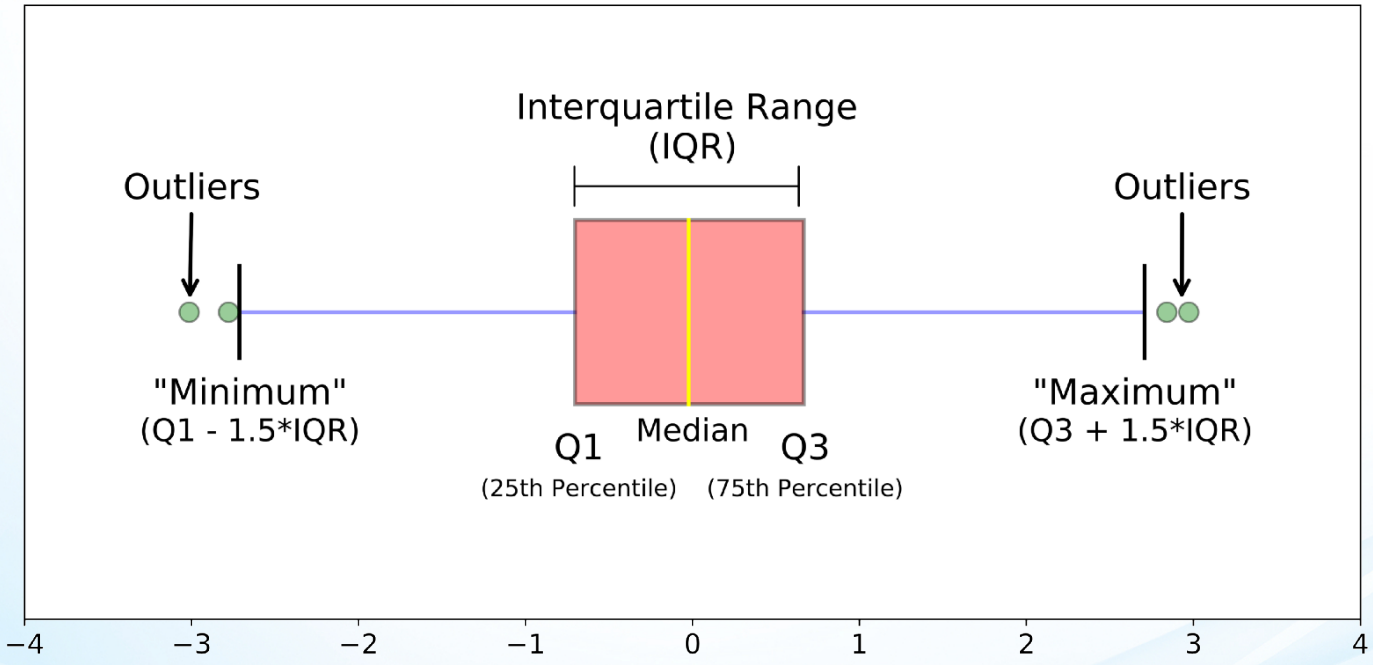
- Основанные на статистических распределениях;
- Основанные на ближайших соседях;
- Случайные леса;
- Одноклассовая классификация (One Class SVM).
- Автоэнкодер

Детектирование как выбросов, так и новизны.

Детектирование аномалий

Статистические методы

Ящик с усами



В гауссовском случае расстояние между хвостами соответствует 99.3% данных, что аналогично правилу трех сигм.

Критерий Граббса

Пусть $X_1, ..., X_n$ — выборка из нормального распределения.

H_0: в выборке нет выбросов

H_1: в выборке есть выбросы

Z-статистика: $Z_i = \left|X_i - \overline{X}\right| / S$,

где S^2 — выборочная дисперсия (несмещенная)

Если

$$\max_{i=1..n} Z_i > \frac{n-1}{\sqrt{n}} \sqrt{\frac{T^2_{n-2,\alpha/(2n)}}{n-2+T^2_{n-2,\alpha/(2n)}}},$$

то X_i, соответствующий максимальному Z_i считается выбросом и удаляется, а процедура повторяется (МПГ не надо).

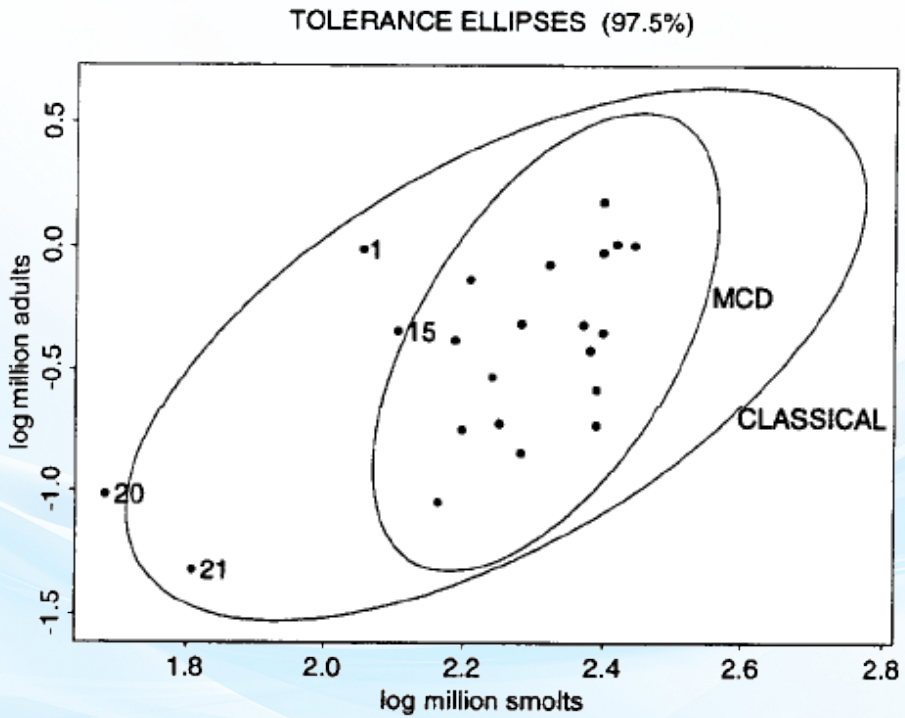
Многомерный случай сводим к одномерному: $Y_i = \sqrt{\left(X_i - \overline{X}\right)^T \widehat{\Sigma}^{-1} \left(X_i - \overline{X}\right)}$

Эллиптическая оболочка (Elliptic Envelope)

$X_1, ..., X_n$ — выборка в d-мерном пространстве.

Предположение: распределение типичных объектов имеет эллипсоидальный вид (например, нормальное).

Идея: найти эллипс, который описывает типичные объекты.



Применение: выбросы, новизна.

Эллиптическая оболочка (Elliptic Envelope)

C-шаг:

Дано: a и Σ — текущие приближ. среднего и матрицы ковариаций.

1. Для каждого объекта вычислить расстояние Махаланобиса

$$d_{a,\Sigma}(i) = \sqrt{(X_i - a)^T \Sigma^{-1} (X_i - a)}$$

2. J — набор из h точек, которые имеют наименьшее $d_{a,\Sigma}(i)$
3. Пересчитать a и Σ по точкам из J.

Примерное описание метода:

1. Рассматривается небольшие случайные подмножества
2. Внутри каждого проводится несколько итераций C-шага
3. Запоминается несколько наилучших приближ. (с малыми $\det \Sigma$)
4. Подмножества объединяются и снова производится несколько итераций C-шага по полученным ранее начальными приближениям.
5. Для наилучших приближ. выполняются C-шаги на всех данных.

Метод главных компонент

$X_1, ..., X_n$ — выборка в D-мерном пространстве

Идея модификации метода:

1. Пусть первые d главных компонент

   описывают бОльшую часть дисперсии данных.

2. Тогда по остальным $D - d$ компонентам

   данные меняются незначительно.

   То есть проекции не сильно отклоняются от нуля.

3. Если проекция точки x на последние $D - d$ главных компонент

   сильно отклоняется от нуля, то это аномалия.

Метод работает лучше для задач детектирования новизны.

Детектирование аномалий

Методы, основанные на ближайших соседях

Локальный уровень выброса (Local Outlier Factor)

Идея: построение локальной “плотности“ точек. Если плотность точки существенно меньше плотности соседей, то это выброс.

$X_1, ..., X_n$ — выборка в метрическом пр-ве с метрикой $\rho(x, y)$.

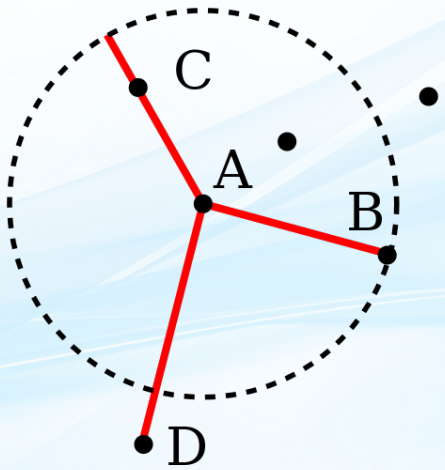
$N_k(X_i)$ — множество k ближайших соседей точки X_i

$\rho_k(X_i)$ — расстояние до k-го соседа точки X_i

Достижимое “расстояние“ из точки X_i:

$$\rho_k^{reach}(X_j|X_i) = \max\{\ \rho_k(X_i),\ \rho(X_i, X_j)\ \}$$

Смысл: все объекты из $N_k(X_i)$ имеют одинаковое “расстояние“ до X_i.



Локальный уровень выброса (Local Outlier Factor)

Достижимое “расстояние“ из точки X_i:

$$\rho_k^{reach}(X_j|X_i) = \max\{\ \rho_k(X_i),\ \rho(X_i, X_j)\ \}$$

Локальная плотность достижимости объекта X_i —

обратное к среднему достижимому расстоянию из $N_k(X_i)$ до X_i

$$ldr(X_i) = \left(\frac{1}{|N_k(X_i)|} \sum_{x \in N_k(X_i)} \rho_k^{reach}(X_i|x) \right)^{-1}$$

Локальный уровень выброса —

отношение средней локальной плотности соседей к лок. плотности X_i

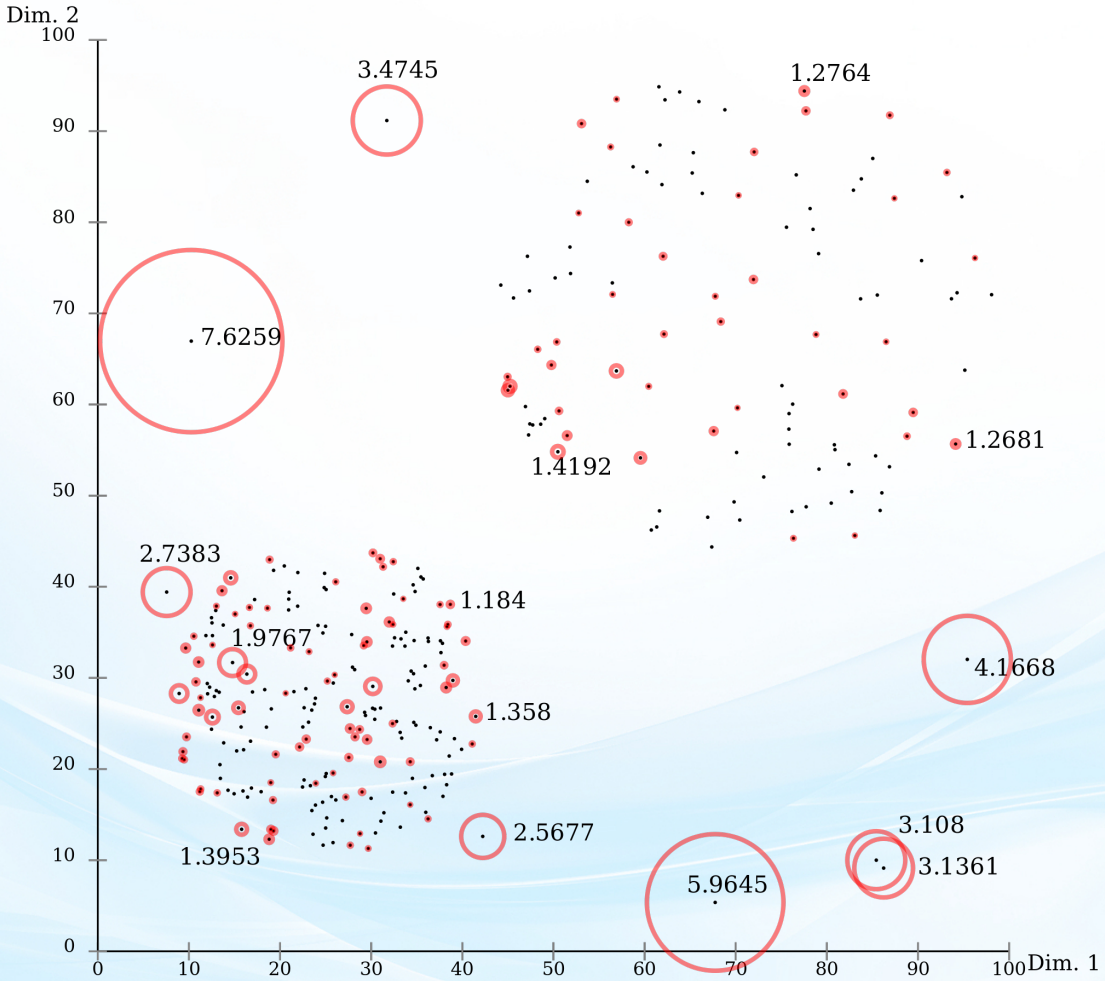
$$LOF(X_i) = \frac{1}{|N_k(X_i)|} \sum_{x \in N_k(X_i)} \frac{ldr(x)}{ldr(X_i)}$$

Локальный уровень выброса (Local Outlier Factor)

$LOF(X_i) \approx 1$ — плотность точки X_i похожа на плотность соседей,

$LOF(X_i) \ll 1$ — точка X_i внутренняя,

$LOF(X_i) \gg 1$ — точка X_i является выбросом,



Применение: выбросы, есть модификация для детектирования новизны.

DBSCAN [напоминание]

DBSCAN = Density-Based Spatial Clustering of Applications with Noise

$X_1, ..., X_n$ — выборка в d-мерном пространстве

Задача: кластеризовать типичные точки и отметить выбросы.

Метод:

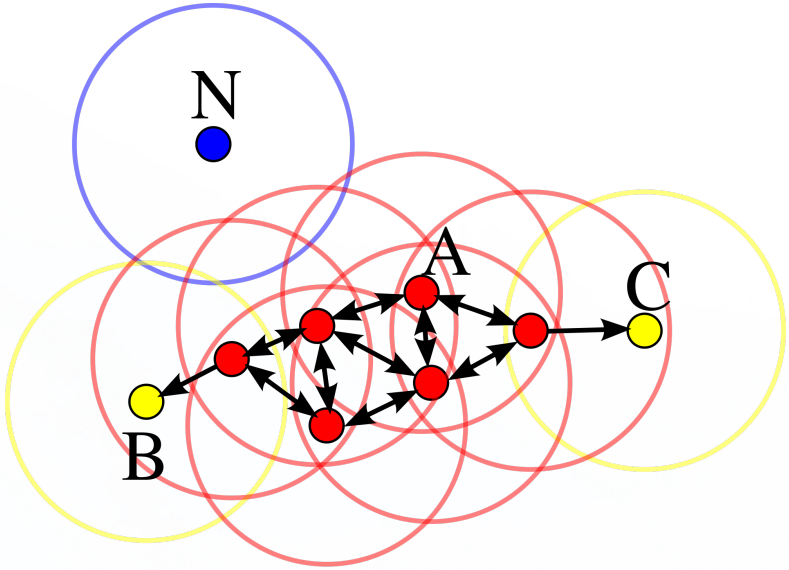
1. Точка x — *основная точка*,

   если в окрестности радиуса ε находится не менее k точек выборки;
2. Точка y *достижима прямо* из основной точки x, если $\|x - y\| \leqslant \varepsilon$
3. Точка y *достижима* из основной точки x,

   если существует путь по основным точкам из x в y
4. Если точка y недостижима из основных точек, то она — выброс.
5. Основная точка формирует кластер

   вместе со всеми достижимыми из нее точками.

DBSCAN



Идея построения:

1. Найти ε-окрестность точки x;
2. Выделить точки с не менее k соседями;
3. Найти связные компоненты среди только основных точек;
4. Назначить неосновную точку ближайшему кластеру,

   если он на расстоянии не более ε. Иначе признать ее выбросом.

Применение: выбросы.

Детектирование аномалий

Случайные леса

Изолирующий лес (Isolation Forest, iForest)

$X_1, ..., X_n$ — выборка в d-мерном пространстве

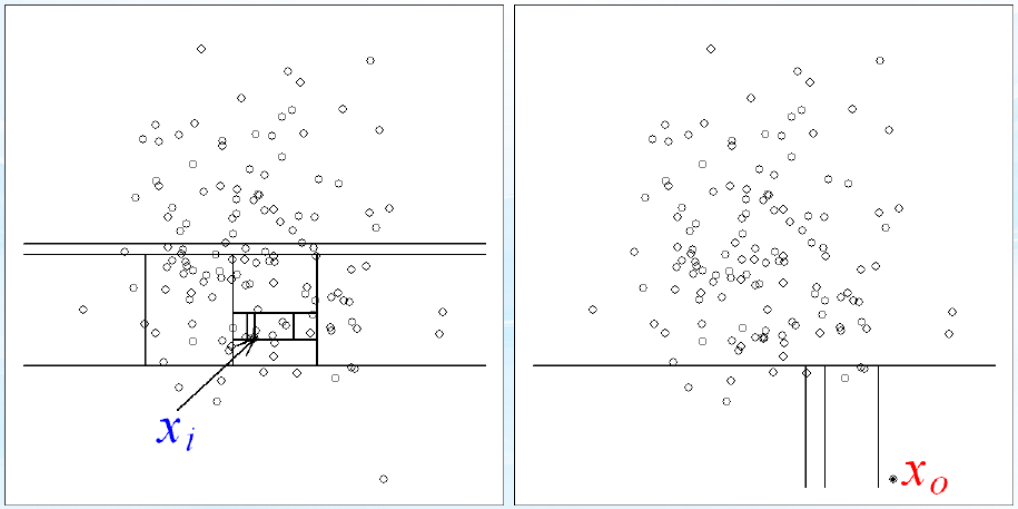
Построение iTree по подвыборке S:

1. Выбираем случайный признак;
2. Выбираем случайное значение x на отрезке $\left[\max\limits_{x \in S} x_j,\ \min\limits_{x \in S} x_j\right]$
3. Делим S по порогу x признака j

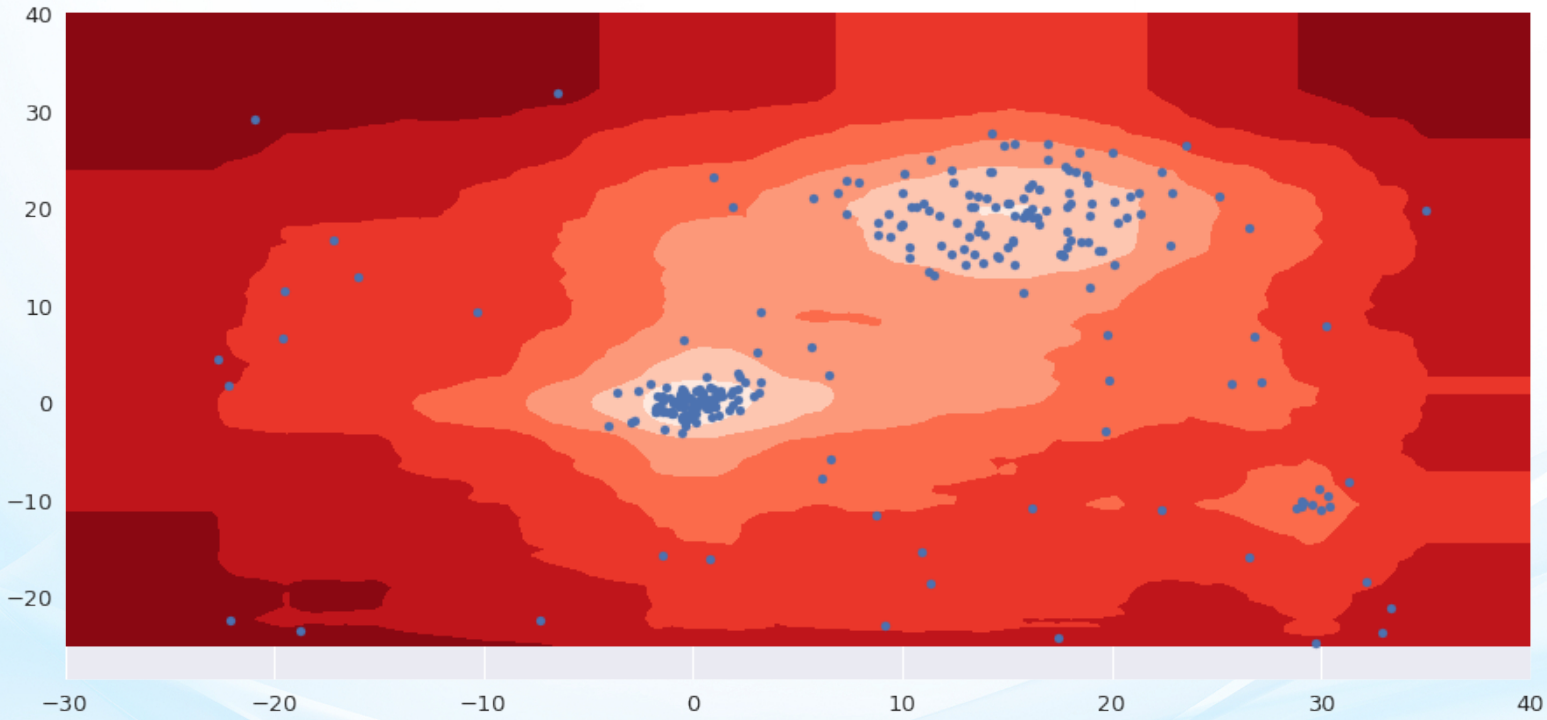
   и рекурсивно строим дерево на полученных частях.

Лес iForest = нескольких независимых iTree.

Мера типичности x = средняя глубина листьев, в которые попал x.



Изолирующий лес (Isolation Forest)



Применение: выбросы, новизна.

Robust Random Cut Forest (Amazon, 2016)

$X_1, ..., X_n$ — выборка в d-мерном пространстве

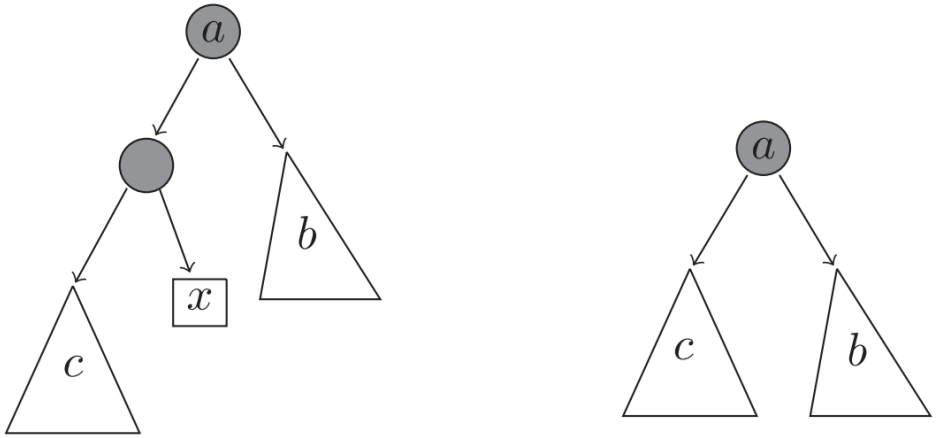
Построение RRCT-дерева $T(S)$ по подвыборке S:

1. Выбираем случайный признак j с вероятностью $r_j \Big/ \sum_{j=1}^{d} r_j$,
   где $r_j = \max\limits_{x \in S} x_j - \min\limits_{x \in S} x_j$
2. Выбираем случайное значение x на отрезке $\left[\max\limits_{x \in S} x_j,\ \min\limits_{x \in S} x_j\right]$
3. Делим S по порогу x признака j
   и рекурсивно строим дерево на полученных частях.

Лес получается построением нескольких независимых деревьев.

Robust Random Cut Forest (Amazon, 2016)

$T(S - \{x\})$ — дерево при удалении объекта x из $T(S)$:



$f(x, T)$ — высота листа объекта x в дереве T.

$M(T(S)) = \sum_{x \in S} f(x, T(S))$ — сложность дерева $T(S)$.

Выбросы находятся вначале $\Longrightarrow$ при удалении меняют высоту многих объектов $\Longrightarrow$ сложность сильно уменьшается.

$$Disp(x, S) = \frac{1}{\#T} \sum_{T} \sum_{y \in S - \{x\}} \Big(f(y, T(S)) - f(y, T(S - \{x\})) \Big)$$

Детектирование аномалий

Одноклассовая классификация

SVM (два класса)

$X_1, ..., X_n$ — точки в d-мерном пространстве

$Y_1, ..., Y_n$ — метки класса $Y_i \in \{-1, +1\}$

$$\begin{cases} \frac{\|\theta\|^2}{2} + C \sum_{i=1}^{n} \xi_i^+ \longrightarrow \min\limits_{\theta, \theta_0, \xi} \\ Y_i \left(\theta^T X_i + \theta_0\right) \geqslant 1 - \xi_i \end{cases}$$

Решение:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \lambda_i Y_i X_i^T x + \theta_0\right),$$

где λ_i из решения двойственной задачи.

SVM (два класса)

$X_1, ..., X_n$ — точки в d-мерном пространстве

$Y_1, ..., Y_n$ — метки класса $Y_i \in \{-1, +1\}$

$$\begin{cases} \frac{K(\theta,\theta)}{2} + C\sum_{i=1}^{n} \xi_i^+ \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i\left(K(\theta, X_i) + \theta_0\right) \geqslant 1 - \xi_i \end{cases}$$

Решение:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i Y_i K(X_i, x) + \theta_0\right),$$

где λ_i из решения двойственной задачи.

One class SVM [детектирование новизны]

$X_1, ..., X_n$ — точки в d-мерном пространстве

Идея: отделить все точки от начала координат гиперплоск. $\theta^T x = \rho$.

$$\begin{cases} \frac{\|\theta\|^2}{2} + \frac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i^+ - \rho \longrightarrow \min\limits_{\theta,\xi,\rho} \\ \theta^T X_i \geqslant \rho - \xi_i \end{cases}$$

ν — верхняя границ доли выбросов;

θ — нормаль к гиперплоскости;

ρ — расстояние от начала координат до гиперплоскости.

Решение:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n}\lambda_i X_i^T x - \rho\right),$$

где λ_i из решения двойственной задачи.

One class SVM [детектирование новизны]

$X_1, ..., X_n$ — точки в d-мерном пространстве

Идея: отделить все точки от начала координат

$$\begin{cases} \frac{K(\theta,\theta)}{2} + \frac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i^+ - \rho \longrightarrow \min\limits_{\theta,\xi,\rho} \\ K(\theta, X_i) \geqslant \rho - \xi_i \end{cases}$$

ν — верхняя границ доли выбросов;

θ — нормаль к гиперплоскости;

ρ — расстояние от начала координат до гиперплоскости.

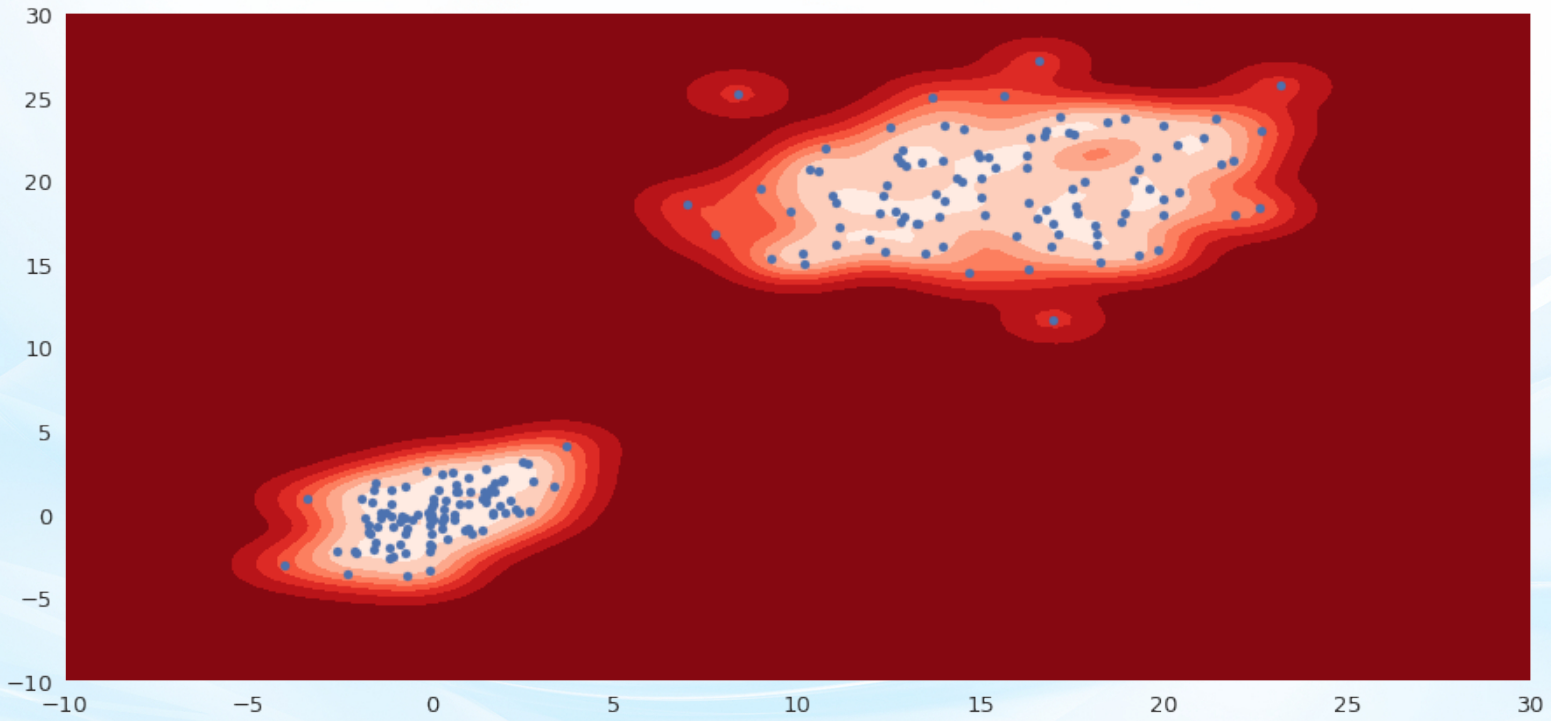
Решение:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n}\lambda_i K(X_i, x) - \rho\right),$$

где λ_i из решения двойственной задачи.

One class SVM [детектирование новизны]

- В настоящее время популярный метод детектирования новизны.
- Хорошо работает только для RBF-ядра.

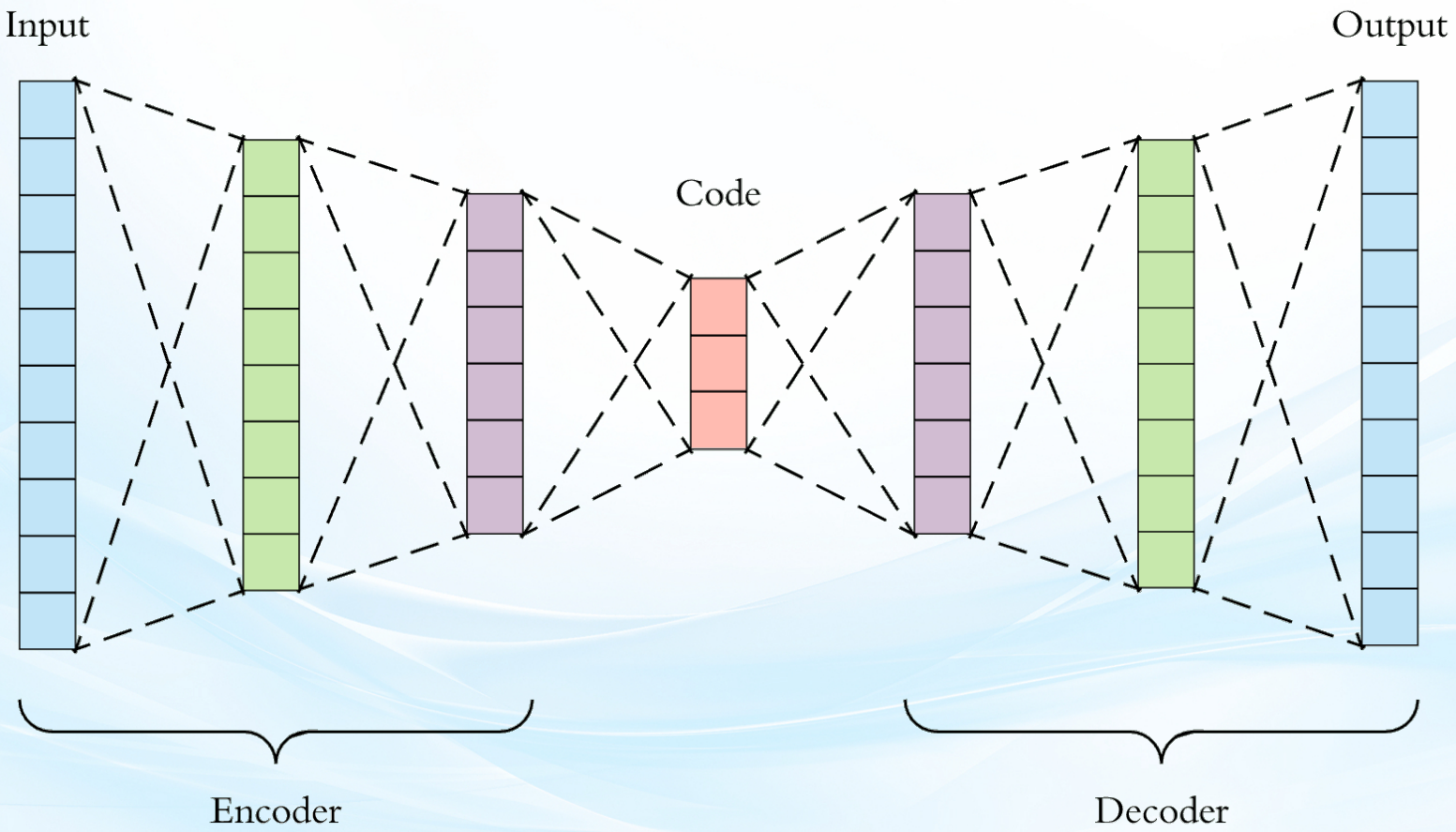


Детектирование аномалий

Автоенкодер

Автоэнкодер [напоминание]

Идея: чем более вероятно событие, тем меньше информации требуется, чтобы его описать.



Автоэнкодер [напоминание]

Особенности модели

- Вход и выход модели совпадают.
- Слой между энкодером и декодером меньше чем входные данные.
- Энкодер и декодор не обязательно полностью симметричны.
- В качестве энкодера и декодера может быть полносвязная сеть, реккурентная, сверточная и т.д., а также их комбинации.

Автоэнкодер: детектирование аномалий

Основные идеи

- За счет сокращения описания данных нетипичное их поведение стирается.
- Для большего эффекта при обучении можно добавить на вход некоторое количество искусственных аномалий. Выход оставить прежним.


ВСЁ!