

Лекция 16

Одноклассовые методы и обнаружение аномалий

Е. А. Соколов
ФКН ВШЭ

8 марта 2021 г.

В задачах кластеризации, о которых шла речь ранее, требуется разделить выборку на группы так, чтобы внутри каждой группы объекты были похожи друг на друга. Теперь мы изучим немного другую постановку — поиск аномалий. В ней даётся выборка «нормальных» объектов, и требуется построить некоторую модель, описывающую данную выборку. Далее для новых объектов требуется определять, принадлежат ли они тому же распределению, что и эта выборка, или же являются выбросами или аномалиями. Такие методы применяются, например, в задачах обнаружения мошеннического поведения или раннего обнаружения неполадок оборудования.

1 Несбалансированная классификация

В некоторых задачах примеры аномалий могут быть даны, но в небольших объёмах — например, при анализе данных систем самолёта может быть известно несколько аномальных ситуаций из прошлого. Такую задачу можно рассматривать как классификацию с несбалансированными классами. При решении обычными методами классификатору оказаться выгоднее относить все объекты к одному классу, поэтому имеет смысл модифицировать процедуру обучения.

Самые простые методы борьбы с несбалансированностью — *undersampling* и *oversampling*. Первый из них удаляет случайные объекты доминирующего класса до тех пор, пока соотношение классов не станет приемлемым; второй дублирует случайные объекты минорного класса. Оптимальное число объектов для удаления или дублирования следует подбирать с помощью кросс-валидации. Отметим, что данные методы применяются лишь к обучающей выборке, а контрольная выборка остается без изменений.

Более сложный метод SMOTE [1] заключается в дополнении минорного класса синтетическими объектами. Генерация нового объекта производится следующим образом. Выбирается случайный объект x_1 минорного класса, для него выделяются k ближайших соседей из этого же класса (k — настраиваемый параметр), из этих соседей выбирается один случайный x_2 . Новый объект вычисляется как точка на отрезке между x_1 и x_2 : $\alpha x_1 + (1 - \alpha)x_2$, для случайного $\alpha \in (0, 1)$.

2 Одноклассовая классификация

Ниже мы будем обсуждать обнаружение точечных аномалий — объектов, которые существенно отличаются от заданной выборки. При этом выделяют и другие типы. Так, контекстными аномалиями называют наблюдения, отличающиеся от наблюдений, близких по некоторому параметру. Например, температура -10° является нормальной в январе, но аномальной в июне.

Когда мы изучали методы обучения с учителем, в начале мы формулировали функции потерь, затем выбирали модели для решения задачи, и, наконец, обсуждали, как оптимизировать параметры этих моделей с точки зрения функции потерь. В случае с поиском аномалий достаточно сложно ввести универсальную функцию потерь — в каждой задаче под аномалией может пониматься что-то своё. Поэтому мы построим обсуждение несколько иначе: будем сразу формулировать методы, которые кажутся разумными, а затем будем выяснять, что именно эти методы считают аномалиями.

§2.1 Статистические методы

В статистических методах предлагается восстановить плотность выборки $p(x)$, и затем определять аномальность объекта на основе того, насколько вероятно его получить из данной плотности. Например, это можно делать через сравнение значения плотности с порогом $[p(x) < d]$ (порог может подбираться, если известно некоторое количество примеров аномалий) или с помощью статистических тестов. Разберём два подхода к восстановлению плотности: параметрический и непараметрический.

2.1.1 Непараметрический подход

Начнём с одномерных величин. Согласно одному из определений неотрицательная функция $p(x)$ является плотностью распределения случайной величины ξ , если её значение в каждой точке равно пределу

$$p(x) = \lim_{h \rightarrow 0} \frac{1}{2h} \mathbb{P}(\xi \in [x - h, x + h]).$$

Воспользуемся этим определением и построим эмпирическую оценку плотности:

$$\hat{p}(x) = \frac{1}{2h} \frac{1}{\ell} \sum_{i=1}^{\ell} [|x - x_i| < h] = \frac{1}{\ell h} \sum_{i=1}^{\ell} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right],$$

где h — ширина окна, регулирующая гладкость эмпирической плотности. Чем больше объектов обучающей выборки в окрестности точки, тем выше будет плотность.

В указанной оценке используется индикатор, что приводит к отсутствию гладкости. Чтобы устранить это, заменим индикатор того, что расстояние меньше ширины окна, на некоторую гладкую функцию $K(z)$:

$$\hat{p}(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right).$$

Здесь $K(z)$ — ядро (не путайте с ядрами Мерсера!), которое должно удовлетворять четырём требованиям:

- чётность: $K(-z) = K(z)$;
- нормированность: $\int K(z)dz = 1$;
- неотрицательность: $K(z) \geq 0$;
- невозрастание при $z > 0$.

Примером может служить гауссово ядро $K(z) = (2\pi)^{-1/2} \exp(-0.5z^2)$.

Оценку плотности легко обобщить на многомерный случай, заменив разность $|x - x_i|$ на некоторую метрику $\rho(x, x_i)$:

$$\hat{p}(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right), \quad (2.1)$$

где $V(h) = \int K\left(\frac{\rho(x, x_i)}{h}\right) dx$ — нормировочная константа. Следует помнить, что число объектов, необходимое для качественной оценки плотности, растёт экспоненциально по мере роста числа признаков. Из-за этого непараметрические методы подходят только для обнаружения аномалий в маломерных пространствах.

2.1.2 Параметрический подход

Параметрический подход состоит в приближении плотности с помощью распределения $p(x | \theta)$ из некоторого семейства $\{p(x | \theta) | \theta \in \Theta\}$ с помощью метода максимального правдоподобия:

$$\sum_{i=1}^{\ell} \log p(x_i | \theta) \rightarrow \max_{\theta}$$

В качестве распределений могут выступать, например, нормальные или смеси нормальных. В пространствах большой размерности может иметь смысл наивное байесовское предположение, о котором пойдёт речь на семинарах.

2.1.3 Обсуждение подхода

Если говорить о параметрическом подходе, то главная проблема — необходимость выбрать семейство распределений. Вряд ли многомерные реальные данные можно описать даже смесью каких-либо стандартных распределений. В случае же с непараметрическим подходом проблема состоит в объёме данных — не факт, что их хватит, чтобы описать выборку достаточно хорошо. Скорее всего, статистический подход будет хорошо только для данных низкой размерности или в случае, когда мы по каким-то причинам точно знаем тип распределения наших данных.

§2.2 Метрические методы

Интуитивно хочется определить аномалию как объект, который не похож на другие объекты нашей выборки. Попытаемся придумать метод, который опирается именно на расстояния между объектами.

Простейший подход основан на выделении объектов, которые расположены от других существенно дальше, чем объекты в среднем удалены друг от друга. А именно, объект x объявим аномальным, если p или меньше процентов объектов имеют до него расстояние меньше ε :

$$\frac{1}{\ell} \sum_{i=1}^{\ell} [\rho(x, x_i) < \varepsilon] \leq p.$$

Пороги p и ε являются параметрами, которые должны настраиваться по известным примерам аномалий или исходя из априорных предположений.

§2.3 Одноклассовый метод опорных векторов

Заметим, что похожим образом можно применять любую модель для обнаружения аномалий — достаточно обучить её так, чтобы прогнозы для объектов из обучения были близки к нулю или, наоборот, как можно сильнее отделены от нуля.

Выше мы пытались описать данные с помощью распределения или использовать метрику, чтобы оценить аномальность объекта. Далее мы разберём подход на основе моделей. Действительно, можно взять любую модель машинного обучения и настроить её так, чтобы на нормальных объектах она принимала близкие к нулю или, например, положительные значения. Тогда можно будет считать, что если на новом объекте прогноз сильно отличается от прогнозов на обучающей выборке, то этот объект скорее аномальный. Мы поговорим о двух методах: на основе SVM и на основе решающих деревьев.

Для обнаружения аномалий, по сути, необходимо построить некоторую функцию $a(x)$, которая принимает значение 1 на области как можно меньшего объёма, содержащей как можно больше объектов выборки; во всех остальных точках она должна иметь значение 0. Такая функция будет компактно описывать обучающую выборку, и можно рассчитывать, что на аномальных объектах она будет отрицательной.

Будем строить линейную функцию $a(x) = \text{sign}\langle w, x \rangle$, и потребуем, чтобы она отделяла выборку от начала координат с максимальным отступом. Соответствующая оптимизационная задача будет иметь вид [2]

$$\begin{cases} \frac{1}{2} \|w\|^2 + \frac{1}{\nu \ell} \sum_{i=1}^{\ell} \xi_i - \rho \rightarrow \min_{w, \xi, \rho} \\ \langle w, x_i \rangle \geq \rho - \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i \geq 0, \quad i = 1, \dots, \ell. \end{cases}$$

Здесь гиперпараметр ν отвечает за корректность на обучающей выборке — можно показать, что он является верхней границей на число аномалий (объектов выборки, на которых $a(x) = -1$). Решающее правило будет иметь вид

$$a(x) = \text{sign}(\langle w, x \rangle - \rho),$$

где ответ -1 будет соответствовать выбросу. Получается, что мы ищем гиперплоскость так, что:

- она отделяет как можно больше объектов выборки от нуля (чем меньше ν , тем больше объектов мы будем отделять) — за это отвечает слагаемое $\frac{1}{\nu\ell} \sum_{i=1}^{\ell} \xi_i$ в функционале;
- она имеет большой отступ $\frac{1}{\|w\|^2}$;
- она при этом как можно сильнее отдалена от нуля (то есть ρ как можно большее значение).

Для данной задачи можно выписать двойственную и сделать ядровой переход в ней:

$$\begin{cases} \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j K(x_i, x_j) \rightarrow \min_{\lambda} \\ 0 \leq \lambda_i \leq \frac{1}{\nu\ell}, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell} \lambda_i = 1. \end{cases}$$

Модель при этом будет иметь вид

$$a(x) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i K(x, x_i) - \rho \right).$$

Заметим, что при использовании гауссова ядра данная модель будет очень похожа на метод, который строит непараметрическую оценку плотности (2.1) с гауссовым ядром и сравнивает её значение с порогом ρ .

При использовании подходящих ядер можно действительно получить функцию, которая точно описывает обучающую выборку в исходном пространстве. Также можно показать, что объекты из того же распределения, из которого сгенерирована обучающая выборка, будут с не очень большой вероятностью попадать в область с отрицательным значением $a(x)$.

§2.4 Isolation forest

Ранее мы обсуждали, что случайный лес вводит функцию расстояния — чем чаще два объекта попадают в один лист, тем более похожими их можно считать. Похожий подход можно использовать и для обнаружения аномалий. Метод, который мы разберём, называют изоляционным лесом (Isolation forest) [3].

На этапе обучения будем строить лес, состоящий из N деревьев. Каждое дерево будем строить стандартным жадным алгоритмом, но при этом признак и порог будем выбирать случайно. Строить дерево будем до тех пор, пока в вершине не окажется ровно один объект, либо пока не будет достигнута максимальная высота. Высоту дерева можно ограничить величиной $\log_2 \ell$.

Метод основан на предположении о том, что чем сильнее объект отличается от большинства, тем быстрее он будет отделён от основной выборки с помощью случайных разбиений. Соответственно, выбросами будем считать те объекты, которые оказались на небольшой глубине.

Чтобы вычислить оценку аномальности объекта x , найдём расстояние от соответствующего ему листа до корня в каждом дереве. Если лист, в котором оказался объект, содержит только его, то в качестве оценки $h_n(x)$ от данного n -го дерева будем брать самую глубину k ; если же в листе оказалось m объектов, то в качестве оценки возьмём величину $h_n(x) = k + c(m)$. Здесь $c(m)$ — средняя длина пути от корня до листа в бинарном дереве поиска, которая вычисляется по формуле

$$c(m) = 2H(m-1) - 2\frac{m-1}{m},$$

а $H(i) \approx \ln(i) + 0.5772156649$ — i -е гармоническое число. Оценку аномальности вычислим на основе средней глубины, нормированной на среднюю длину пути в дереве, построенном на выборке размера ℓ :

$$a(x) = 2^{-\frac{\frac{1}{N} \sum_{n=1}^N h_n(x)}{c(\ell)}}.$$

Для ускорения работы можно строить каждое дерево на подвыборке размера s ; в этом случае во всех формулах выше нужно заменить ℓ на s .

Список литературы

- [1] *Chawla N., Bowyer K., Hall L., Kegelmeyer W.* (2002). SMOTE: Synthetic Minority Over-sampling Technique. // Journal of Artificial Intelligence Research, Vol. 16, Pp. 321–357.
- [2] *Schölkopf, Bernhard and Williamson, Robert and Smola, Alex and Shawe-Taylor, John and Platt, John* (1999). Support Vector Method for Novelty Detection. // NIPS'99.
- [3] *Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua* (2008). Isolation forest. // Data Mining, 2008. ICDM'08.