Методы современной прикладной статистики
4.1. Выбросы

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Предварительные замечания

Выброс – результат измерения, выделяющийся из общей выборки, являющийся ошибкой сбора или обработки данных.

Удалять выбросы можно только в том случае, если вы абсолютно уверены, что они не являются значимыми для последующих статистических выводов о данных.

Предварительные замечания

Зачастую выбросы не удаляются из данных, а учитываются при подсчете статистик с меньшим весом, чем остальные наблюдения, как в методах робастной и непараметрической регрессии, или вовсе не учитываются, как в случае использования выборочной медианы.

Удалять выбросы и после этого рассматривать на очищенных данных параметрическую модель не рекомендуется, потому что выброшенные данные могут быть значимыми для модели.

Предварительные замечания

Выбросы можно распознать визуально (с помощью методов понижения размерности понизив размерность данных до 2), а также с помощью методов кластеризации (если в кластере один объект, то велика вероятность, что это выброс).

Методы машинного обучения для борьбы с выбросами: one-class SVM, robust covariance (расстояние Махаланобиса), isolation forest, local outlier factor.

Предварительные замечания

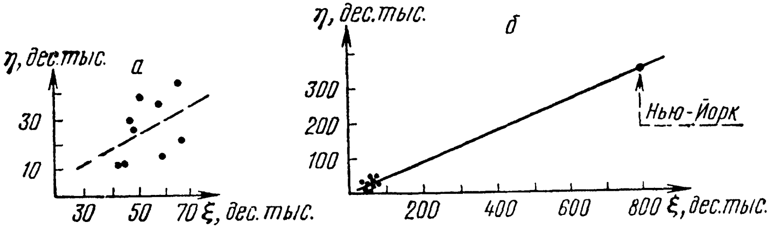


Рис. 1.1. Корреляционное поле, характеризующее связь между численностью населения ξ и числом установленных телевизионных точек η в США в 1953 г.:
а) в девяти городах; б) в десяти городах

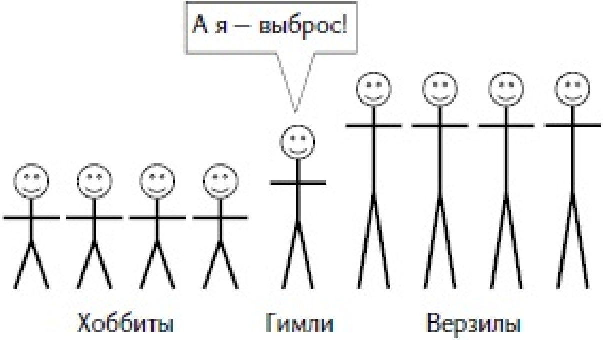
R^2 первой модели равняется 0.18, второй – 0.99. Стоит ли считать Нью-Йорк выбросом в данной модели?

Метод Тьюки работает только для одномерных данных и используется при построении ящика с усами. Если элемент выборки не попадает в интервал

$$(Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)),$$

то он объявляется выбросом. Здесь Q_1 и Q_3 – выборочные квантили уровня 0.25 и 0.75 соответственно, параметр k, как правило, полагается равным 1.5.

Границы Тьюки адекватно работают, если плотность распределения имеет колоколообразный вид и выброс находится не в середине выборки, а с краю от неё.



Пусть $X_1, \ldots, X_n$ – выборка и X_i – некое отстоящее от выборки наблюдение. Если

$$\left|\frac{X_i - \overline{X}}{s}\right| > u_{1-\frac{1}{4n}},$$

то считаем, что X_i является выбросом. Здесь $u_{1-\frac{1}{4n}}$ – квантиль $N(0, 1)$ соответствующего уровня, $\overline{X}$ – выборочное среднее, s – корень из выборочной дисперсии.

Аналогично границам Тьюки, критерий не различает выбросы, находящиеся по центру значений выборки.

Предполагаем, что выборка $X_1, \ldots, X_n$ сделана из нормального закона. Рассмотрим статистику Граббса

$$G = \frac{\max_i |X_i - \overline{X}|}{s},$$

где $\overline{X}$, как и ранее, выборочное среднее, s – корень из выборочной дисперсии.

Обозначим u – квантиль уровня $\frac{\alpha}{2n}$ распределения Стьюдента с $n-2$ степенями свободы. Тогда если

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{u_\alpha^2}{n-2+u_\alpha^2}},$$

то отвергаем гипотезу об отсутствии выбросов на уровне значимости α.

В отличие от критерия Шовенэ и Q-критерия Диксона, критерий Граббса является настоящим статистическим критерием.

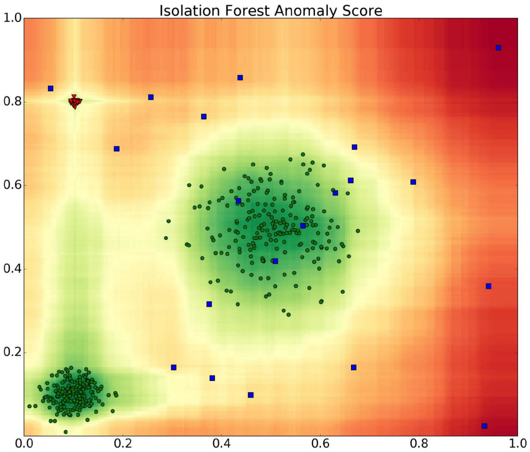
Кроме того, критерий Граббса можно применить и для удаления нескольких выбросов. Для этого после удаления очередного выброса нужно пересчитывать статистику Граббса и критический уровень с учетом понижения n на единицу.

Isolation forest

Алгоритм Isolation forest заключается в построении случайного бинарного решающего дерева. В очередном узле выбирается случайный признак и случайный порог разбиения (выбирается равномерно на отрезке от минимального до максимального значения признака).

Когда число листьев совпадет с числом объектов, алгоритм останавливается. Ответом на листе объявляется его глубина в построенном дереве.

Утверждается, что аномальным точкам/выбросам свойственно оказываться в листьях с низкой глубиной. Окончательный результат усредняется по нескольким запускам алгоритма.



Пример работы алгоритма Isolation forest. Красным цветом отмечена низкая глубина листа соответствующего объекта, зеленым - высокая.

Алгоритм обладает рядом несомненных преимуществ:

- распознает аномалии разных типов: как изолированные точки, так и кластеры аномалий малых размеров;
- не требует больших затрат по памяти, в отличие от многих метрических алгоритмов, где часто требуется хранить матрицы попарных расстояний;
- обладает только одним параметром, требующим подбора - процент данных, которые нужно отбросить;
- инвариантен к масштабированию признаков;
- устойчив к увеличению размерности пространства признаков.

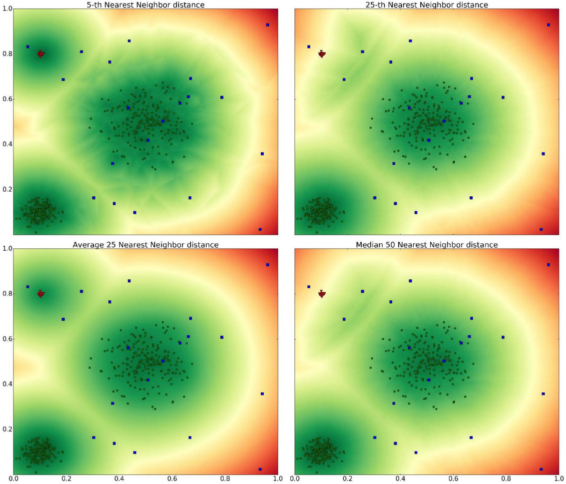
Метрические методы

Пусть на пространстве признаков задана метрика $\rho(x, y)$.
В качестве anomaly score можно брать

- расстояние до k-того ближайшего соседа или среднее/медиану расстояний до k ближайших соседей;
- значение локальной плотности, заданной непараметрически (например, с помощью ядерных оценок);
- размер кластера, в котором оказалась точка.

Достоинства: результаты легко интерпретируемы.
Недостатки: непонятно, как выбирать k и метрику; даже при наилучшем подборе метрики аппроксимация может оказаться грубой.

Примеры работы метрических алгоритмов.

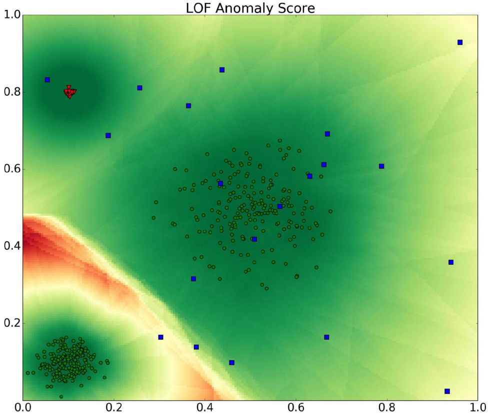


Но локальная плотность точки, лежащей в небольшом кластере аномалий, может оказаться выше, чем для любой точки из большого кластера нормальных данных. Эту проблему призван исправить метод Local outlier factor.

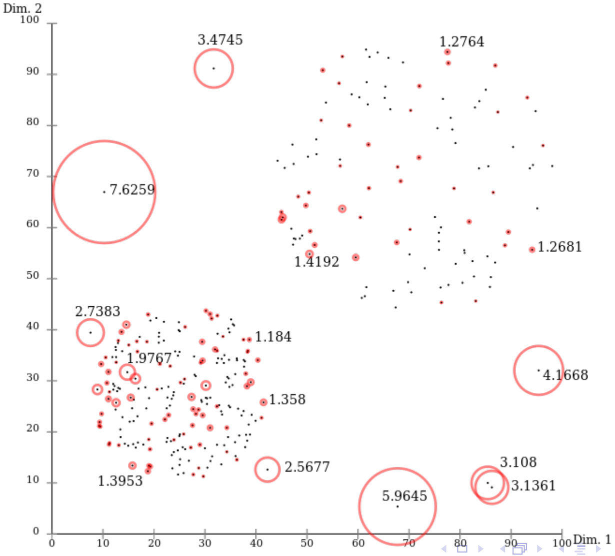
Определим $D_k(y)$ – расстояние от точки y до k-того ближайшего соседа. Досягаемостью точки x относительно y называется величина $R_k(x, y) = \max(\rho(x, y), D_k(y))$.

Пусть $AR_k(x)$ – средняя досягаемость x относительно $N_k(x)$, множества k своих ближайших соседей, тогда

$$LOF_k(x) = \frac{1}{k} \sum_{y \in N_k(x)} \frac{AR_k(x)}{AR_k(y)}.$$



Значения $LOF \approx 1$ говорят о том, что объект находится в кластере, значительно больше 1 – что объект изолирован, меньше единицы – что объект в области локально высокой плотности.


Dim. 2
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Dim. 1
3.4745
1.2764
7.6259
1.4192
1.2681
2.7383
1.184
1.9767
4.1668
1.358
2.5677
1.3953
5.9645
3.108
3.1361

Finita!