Методы обнаружения выбросов

1) Поиск аномальных объектов по общему смыслу

Известен нормальный диапазон для значений признака.
Пример выброса: человек с ростом более 200см

2) Методы, основанные на анализе одного признака.

3) Методы, основанные на одновременном анализе нескольких признаков.

Методы, основанные на анализе одного признака

Имеем следующие значения признака: $P = (p_1, \dots, p_n)$

Нужно найти значения p_i , расположенные вдали от среднего значения.
Пусть $\bar{p}$ - среднее значение, *n* – объём выборки, *S* – отклонение

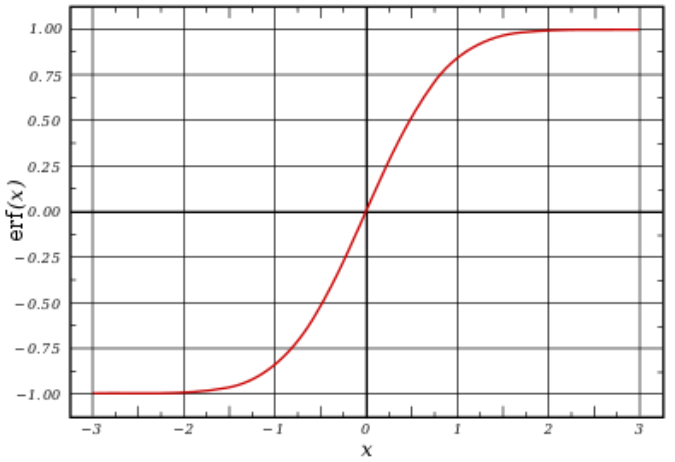
Метод 1. Удалить все объекты, если величина $|p_i - \bar{p}| > eps$

Метод 2. Удалить все объекты, если величина $\frac{|p_i - \bar{p}|}{S} > eps$

Метод 3. Критерий Шовене (Chauvenet)
Значение p_i является выбросом, если выполнено неравенство:

$$\operatorname{erfc}\left(\frac{|p_i - \bar{p}|}{S}\right) > \frac{1}{2n},$$

где функция ошибок $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$, а её дополнение $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$



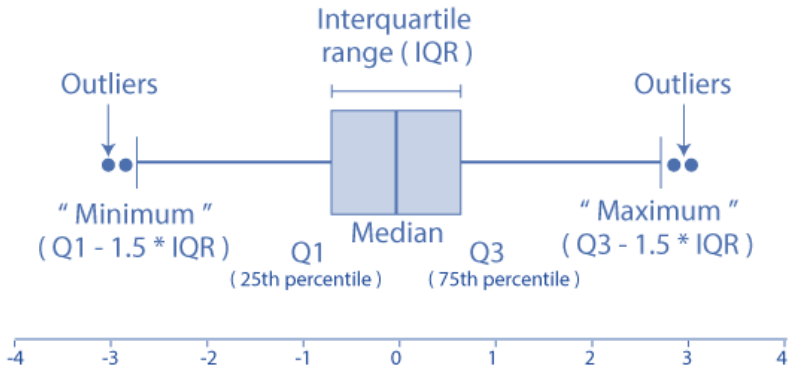
P.S. Значения среднего и отклонения сильно чувствительны к наличию выбросов.

Метод 4. «Ящик с усами» или метод, основанный на квартилях

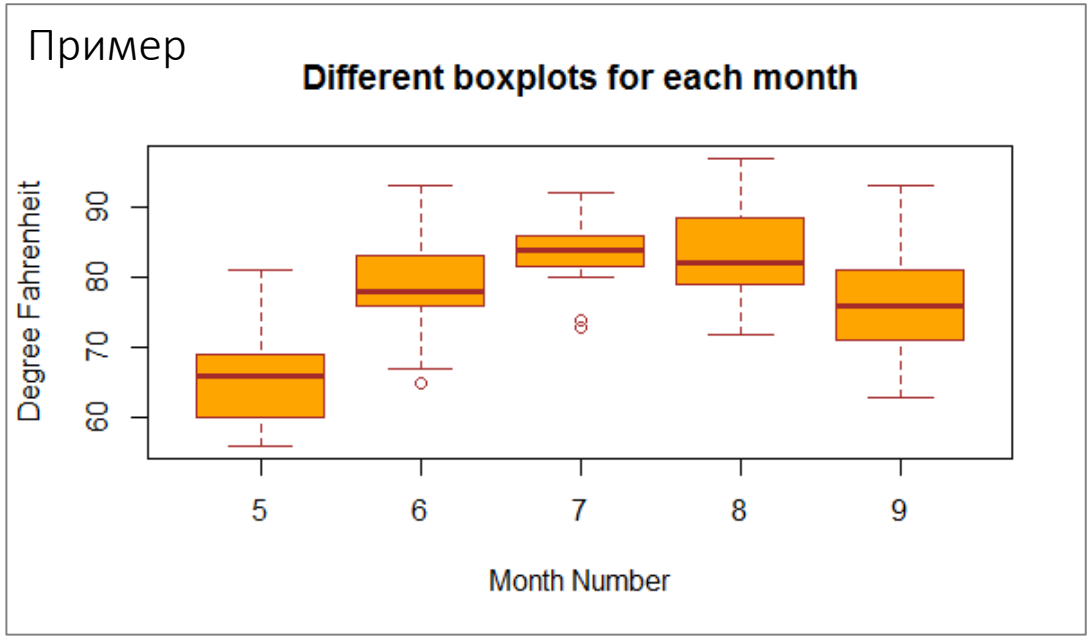
1 квартиль Q25: это такое число, что ровно 25% выборки меньше его.
2 квартиль Q50: это такое число, что ровно 50% выборки меньше его(медиана).
3 квартиль Q75: это такое число, что ровно 75% выборки меньше его.

Упражнение: Для выборки (-11, 1, 2, 3, 4, 5, 5, 6, 7, 500)
Q25=? Q50=? Q75=?

Если элемент не попадает в интервал *(Q25-1.5*(Q75-Q25), Q75+1.5*(Q75-Q25))*, то элемент – выброс.
Пример:
В примере (-5, 0, 1, 2, 4, 5, 5, 6, 7, 100)
Q25=1, Q75=6
Интервал: (1-1.5*5, 6+1.5*5)= (-6.5,13.5) -> 100 – выброс.



Пример



Методы, основанные на анализе нескольких признаков

Метод 1. Метрические методы
Найти расстояние от каждого объекта до его ближайшего соседа.
У выбросов такое расстояние будет большим

Метод 2. Поиск выбросов с помощью кластеризации
Запустить алгоритм кластеризации. Он разобьет объекты на группы.
Выбросы – это элементы малых (в том числе и одноэлементных) групп.

Метод 3.
Вычислить выпуклую оболочку объектов
(как точек в m-мерном пространстве).
Выбросами будут объекты на границе.

Метод 4.
Поиск выбросов с помощью моделей
предсказания(SVM).