MSMS 302 - Outlier Detection

Ananda Biswas

An outlier is a point that is far away from most of the data points. Normal points exist in dense neighbourhood, while outliers are isolated. The study of outlier detection is helpful in fraud detection, trojan detection, cyber-security, fault detection in manufacturing, healthcare, identifying novel pattern in scientific data.

- Types of Outliers :

1. Global / Point Outlier : A point far away from the rest of the data is global outlier.
2. Contextual Outlier (Conditional Outlier) : Outlier in specific context; temperature 30° is normal in Delhi but outlier in Ladakh.
3. Collective Outlier : A group of data points that are unusual together. If they are studied isolated, they may not be outliers. For example, sudden spike in server activity during midnight.

There are two types of methods for outline detection.

(i) Density based method : In this context, an outlier is a point in a region of low density compared to its neighbourhood. Unlike the other Distance based method, Density based method considers local variation in density.

A formal definition : Given a data-set $X = \{\underset{\sim}{x_1}, \underset{\sim}{x_2}, \ldots, \underset{\sim}{x_n}\}$, $\underset{\sim}{x_i} \in \mathbb{R}^d$, a certain point $\underset{\sim}{x_i}$ will be outlier if $P(\underset{\sim}{x_i}) < \epsilon$ where ϵ is a threshold. There are several density based methods. Some are given below.

(a) Statistical method : Assume data points follow normal distribution. Then for a point $x_i \in \mathbb{R}$, calculate $z_i = \dfrac{x_i - \mu}{\sigma}$. If $|z_i| > 3$, we often consider x_i as outlier.

(b) Local Outlier Factor (LOF) : LOF measures how isolated a point is related to its neighbourhoods. In this method,

- Step (1) : For each point p, compute k-distance(p) = distance of p from its k-th nearest neighbour.
- Step (2) : For each point p, find the set of k points closest to it.
- Step (3) : For each pair of points p, o, calculate reachability distance as

$$\text{reach-dist}_k(p, o) = max\{k\text{-distance}(p), d(p, o)\}$$

- Step (4) : For each point p, calculate Local Reachability Density as

$$\text{LRD}_k(p) = \frac{1}{\dfrac{1}{|N_k(p)|}\displaystyle\sum_{o \in N_k(p)} \text{reach-dist}_k(p, o)}.$$

- Step (5) : Compute LOF Score as

$$\text{LOF}_k(p) = \frac{\displaystyle\sum_{o \in N_k(p)} \frac{\text{LRD}_k(o)}{\text{LRD}_k(p)}}{|N_k(p)|}.$$

If LOF ≈ 1, then the point is normal.

If LOF > 1, the point more likely to be an outlier.

(c) Mahalonobis Distance : In multivariate case, if the data are assumed to be coming from Multivariate Normal distribution, then

$$D(\underset{\sim}{x}) = \sqrt{(\underset{\sim}{x} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu})} \sim \chi^2_d.$$

If $D(\underset{\sim}{x}) > \lambda$, λ being a threshold obtained from χ^2_d distribution, $\underset{\sim}{x}$ can be considered as an outlier.

(ii) Distance based methods : For a given data-set $X = \{\underset{\sim}{x_1}, \underset{\sim}{x_2}, \ldots, \underset{\sim}{x_n}\}$, $\underset{\sim}{x_i} \in \mathbb{R}^d$, have a distance function $d(\underset{\sim}{x}, \underset{\sim}{y})$. A point $\underset{\sim}{x}$ is outlier if

$$\left|\{\underset{\sim}{y} \in X | d(\underset{\sim}{x}, \underset{\sim}{y}) \leq r\}\right| < \pi_r$$

where π_r is the minimum fraction of points required at distance r.

(a) kNN method : For each point x compute distance to its k-nearest neighbours denoted as $d_k(x)$. Outlier score $O(x)$ is defined as $O(x) = d_k(x)$. If $O(x)$ is very large, the point is outlier.

(b) ML based methods : One-class SVM, Autoencoders, Isolation Forest, Neural Networks etc.