# Outlier detection

**Definition:** Objects with behaviors that are very different from expectation are called outliers or anomalies.

**Types of Outliers:**

- Global Outliers : A data object is a global outlier if it deviates significantly from the rest of the dataset.
- Contextual Outliers : A data object is a contextual outlier if it deviates significantly from the rest of the dataset with respect to some specific context.
- Collective Outliers : A subset of data objects forms a collective outlier, if the objects as a whole deviate significantly from a dataset.

**Challenges in outlier detection:**

- Modeling normal objects and outliers effectively.
- Application specific outlier detection : Choosing similarity / difference measure and the relationship to describe data objects. Eg: In clinical data, small deviation might be enough to be classified as an outlier.
- Handling Noise : Noise can distort data hiding distinction between normal objects and outliers. Noise or outlier?
- Understandability : Explanation of why the detected objects are outliers?

**Outlier detection methods:**

- Supervised, Semi-supervised and Unsupervised methods.
- Statistical, Proximity based methods.
- Clustering based methods.
- Classification based methods.

1. **Supervised Methods**

   **Approaches:**

- Domain experts examine and label a sample of underlying data. Outlier detection then can be modeled as a classification problem.
- Experts may just only label the normal objects. Any other objects not matching the model of normal objects are outliers.
- Model only the outliers and treat objects not matching the model as normal.

**Challenges:**

- Class imbalance: Population of outliers is significantly smaller than that of normal objects.
- Importance of sensitivity / recall

2. **Unsupervised Methods**

Unsupervised methods make an implicit assumption that the normal objects are somewhat clustered. The idea is to find clusters first and any objects not belonging to any other clusters are detected as outliers. However the issues with this approach are:

- Outlier or noise?
- It is often costly to find clusters first then find outliers.

3. **Semi-supervised Methods**

When some labeled normal objects are available, they can be used together with unlabeled objects that are closed by to train a mode for normal objects. The objects not fitting into this model are classified as outliers.

4. **Statistical methods:**

- Make assumptions of data normality i.e. the normal data objects are generated by a statistical (stochastic) model and the data that do not follow this model are outliers.
- Idea is to learn a generative model fitting the given dataset and then identify the objects in low probability regions of model as outliers.

- Types: Parametric and Non-parametric. In parametric, assumes normal data objects are generated by parametric distributions and in Non-parametric, tries to determine the model from input data.

**Detection of Univariate Outliers:**

1. Based on normal distribution : Assume data follows normal distribution. As μ ± 3σ region contains 99.7% of data under assumption of normal distribution, any object that is more than 3σ away from mean of estimated distribution can be considered outliers.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

2. Grubb's test: aka the maximum normed residual test. We calculate z-score for each data object. Then an object is an outlier if :

$$z \geq \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}},$$

where $t^2_{\alpha/(2n),\,N-2}$ is the value taken by t-distribution at a significance level of α/(2n) and N is the number of objects in the dataset.

**Detection of Multivariate Outliers**

The main idea is to transform multivariate outlier detection task to a univariate outlier detection problem.

1. Using Mahalanobis distance: For an object o in the dataset, Mahalanobis distance from o to mean vector ō is given by:

$$\mathbf{MDist(o,\bar{o}) = (o - \bar{o})^T S^{-1} (o - \bar{o})}$$

where S is the covariance matrix.
Since, MDist(o,ō) is a univariate variable, thus Grubb test can be applied to this

measure.

Steps:

- Calculate the mean vector from the multivariate dataset.
- For eat object o, calculate MDist(o, ō).
- Detect outliers in the transformed univariate set.
- If MDist(o, ō) is determined to be an outlier, then o is regarded outlier as well.

**Using mixture of parametric Distributions**

When the actual data distribution is complex, we assume that the data were generated by a mixture of parametric distributions.

**Using histogram**

Steps:
- Construct histogram using the input data.
- If object falls in one of the bins, it is considered normal, else outlier. Or, assign outlier score to the object. For eg: the score can be the inverse of the volume of the bin in which it falls. Larger the score, more likely an outlier.

5. **Proximity based approach**

Assumption : The proximity of an outlier object to its nearest neighbors deviates significantly from the proximity of the object to most of the other objects in the data set. It is of two types:

- Distance based
- Density based

**Distance based outlier detection**

Consults neighborhood of an object, defined by a given radius. An object is considered an outlier if its neighborhood does not have enough other points.

Let $r \geq 0$ be a distance threshold and $0 < \pi \leq 1$ be a fraction threshold. An object o, is a DB(r,π) outlier if

$$|| o' \mid dist(o, o') \leq r) || / || D || \leq \pi$$

**Density based outlier detection**

Investigates the density of an object and that of its neighbors. An object is an

outlier if its density is significantly lower than that of its neighbors.

Some Terminologies:

- k-distance of an object $dist_k(o)$: Distance between a object and its kth nearest neighbor.
- k-distance neighborhood $N_k(o)$: All objects whose distance to o is not greater than k-distance.
- Reachability distance from o′ to o: $reachdist_k$( (o <— o′) = $max\{dist_k(o), dist(o, o')\}$
- Local outlier factor of an object o: The average ratio of the local reachability density of o and those of o's k-nearest neighbors.

$$LOF_k(\boldsymbol{o}) = \frac{\sum_{\boldsymbol{o'} \in N_k(\boldsymbol{o})} \frac{lrd_k(\boldsymbol{o'})}{lrd_k(\boldsymbol{o})}}{\|N_k(\boldsymbol{o})\|}$$

- $LOF_k(o) \sim 1$ means similar density.
- $LOF_k(o) < 1$ means higher density than neighbors.
- $LOF_k(o) > 1$ means lower density than neighbors.

6. **Clustering based approach**

An object is an outlier if:

- It does not belong to any cluster. : DBSCAN
- There is large distance between the object and the cluster to which it is closest.: Partition using k-means. For each object assign outlier score i.e. dist(o/c)/avg_dist(c). The larger the ratio, more likely it is an outlier.
- It is part of a small or sparse cluster.: Find cluster, sort them in decreasing size and assign cluster-based local outlier factor (CBLOF)

For point in large cluster, CBLOF = Cluster size * similarity between point and

cluster. For point in small cluster, Cluster size * similarity between point and closest largest cluster.

The points with lowest CBLOF scores are suspected outliers.

7. **Classification based approach**

Train a classification model that can distinguish normal data from outliers.
The two methods are:

**One-class method**

A classifier is built to describe only the normal class. Any samples that do not belong to the normal class are regarded as outliers. We can learn boundary of normal class using SVM.

**Semi supervised learning**

- Find a large cluster C and a small cluster C1 such that many objects in C carry label normal and objects in C1 carry label outlier.
- We treat every element in C as normal.
- We use one class model of this cluster to identify normal objects.
- We declare all objects in C1 as outliers.
- Any object that does not fall into the model for c, is considered an outlier as well.

**Mining Contextual Outliers**

Contextual outlier detection can be divided into two categories according to the whether the contexts can be clearly identified.

1. Transforming contextual outlier detection to conventional outlier detection: Identify the context of object using contextual attributes and then calculate outlier score for the object in context using conventional methods.
2. Modeling normal behavior with respect to contexts: Train a model that predicts expected behavior attribute values wrt contextual attribute values. Then apply the model to contextual attributes of the object and predict the behavioral attributes. If behavior attribute of any object

deviate significantly from the values predicted by model, then object is considered as contextual outlier.

**Mining Collective Outliers**

It can also be divided into two categories:

1. Method that reduces the problem to conventional outlier detection.
2. Method that models the expected behavior of the structure units.