# Machine Learning Notes

**by Mattia Orlandi**

# 6. Outlier Detection

## 6.1. Problem Description

Anomaly $\leftrightarrow$ Outlier

Causes of anomalies:

- data from different classes;
- natural variation;
- data measurement and collection errors $\Rightarrow$ data cleaning.

Approaches to Anomaly Detection:

- **Model-based techniques**: build a model of data, and outliers will fit poorly in it.
- **Proximity-based techniques**: objects in low-density regions can be considered outliers.

Use of class labels:

- **Supervised**: training set with both anomalous and normal objects (problem of *imbalanced classes*).
- **Unsupervised**: labels not available $\Rightarrow$ learn from training set a way to assign to each object a score reflecting the degree of anomaly.
- **Semi-supervised**: training set contains only normal objects $\Rightarrow$ compute anomaly score from information available for normal objects (one-class classification).

Issues:

- Number of attributes used:
  - single attributes values can be anomalous;
  - common values can be anomalous when considered together.
- Global or Local Perspective: an object may seem unusual w.r.t. all objects, but usual w.r.t. its neighbours.
- Degree of Anomaly: instead of a binary decision, the degree allows to set a tunable threshold.

- Operation:
  - one-anomaly-at-a-time: find most anomalous object, remove it from data and repeat;
  - many-anomalies-at-once:
    - find a set of anomalous objects;
    - problem of **masking**: similar anomalies can mask each other;
    - problem of **swamping**: anomalies distort data model and thus even normal objects seem anomalous.
- Evaluation: usual measures of evaluation are uneffective due to the unbalancing of normal and anomalous classes.
- Efficiency: classification and statistical methods are expensive to setup but lightweight at runtime, and proximity methods tend to have $\mathcal{O}(N^2)$ complexity.

# 6.2. Statistical approaches

Probabilistic definition: an outlier is an object that has a low probability w.r.t. a probability distribution model of data.

- Probability distribution model is created from data by **estimating parameters** of a **user-specified** distribution.
- Statistical tests to identify discordant observations.

Issues:

- identifying specific distribution;
- number of attributes used;
- mixture of distributions.

In an univariate normal distribution:

- an object with an attribute value $x \sim N(0, 1)$ is an outlier if $|x| \geq c$, where $P(|x|) \geq c = \alpha$;
- $\alpha$ is the probability of a false positive.

Strenghts and weaknesses:

- strong theoretical foundation;
- several methods for outlier detection tests for univariate data;
- fewer methods for multivariate data;
- bad performance with high-dimensional data.

# 6.3. Proximity-based Outlier Detection

An object is anomalous if it's distant from most points $\Rightarrow$ proximity measure.

- For each object, make a sorted list of its neighbours according to proximity.
- The outlier score of an object is its distance to its $k$-nearest neighbour.

- Highly sensitive to the value of $k$:
  - if it's too low, nearby outliers will have a low score;
  - if it's too high, normal objects in low-density clusters will have a high score.
- Alternative definition: given a positive real number $R$ and a positive integer $k$, an object is a *distance-based outlier* if less than $k$ objects lie within distance $R$ from the object itself.

### Proximity-based solutions

Finding the **top $m$ outliers** in a dataset, using the notion of distance of the $k$-th nearest neighbours and brute-force solutions, has a $\mathcal{O}(N^2)$ complexity $\Rightarrow$ more efficient algorithms:

- Bay's algorithm:
  - for each example in $\mathcal{E}$ keep track of the $k$ nearest neighbours found so far;
  - determine the **cutoff** value of the score as the distance of the $k$-th nearest neighbour of the top $m$-th outlier found so far;
  - when an example achieve a score lower than the cutoff it is removed since it cannot be an outlier;
  - later iterations find increasing scores, and pruning is more efficient;
  - if data are in random order the average complexity is *quasi*-linear, whereas in worst case is $\mathcal{O}(N^2)$.

## 6.4. Density-based Outlier Detection

- Outliers are found in low-density areas.
- Density-based definition: the outlier score of an object is the inverse of the density around the object.
- Inverse distance:

$$\text{density}(x, k) = \left( \frac{\sum_{y \in Nb(x,k)} \text{distance}(x, y)}{k} \right)^{-1}$$

  where $Nb(x, k)$ is the set containing the $k$-nearest neighbours of $x$.
- Alternative definition of density around an object: it's the number of objects within a specified distance $d$ from the object (if $d$ is too small the density can be underestimated).
- Average Relative Density: avoids outlier detection problems when data contains regions of different densities $\Rightarrow$ better outlier score.

$$\text{ard}(x, k) = \frac{\text{density}(x, k)}{\sum_{y \in Nb(x,k)} \text{distance}(y, k)/k}$$

Strenghts and weaknesses:

- works well when data has regions of different densities;
- natural complexity of $\mathcal{O}(N^2)$, but it can be reduced to $\mathcal{O}(N \log(N))$ for low-dimensional data with special data structures;

- parameter selection is quite difficult.