VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

# Machine Learning

## Outlier Analysis

Jan Platoš

October 1, 2025

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

# Outlier Analysis

# Outlier analysis

**Informal definition (Hawkins)**
An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.

**Outlier applications**

- Data cleaning
  - An outlier represent a noise data.
- Credit card fraud
  - Credit card activity outside the usual pattern may represent an fraud.
- Network intrusion detection
  - Unusual records in traffic that do not follow the regular patterns.

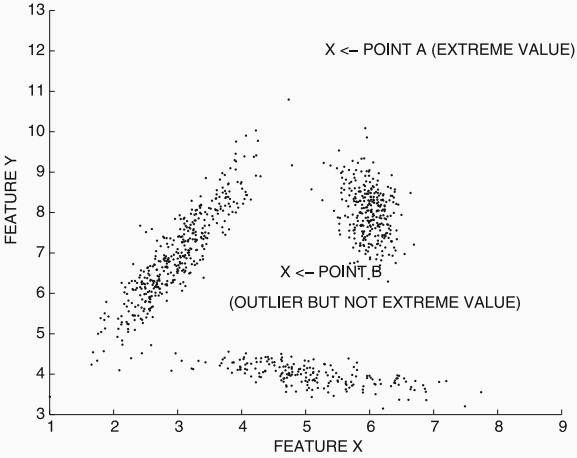
# Outlier analysis - Classification types

- Real-valued outlier score
  - A real value that represents the outlierness of a data point.
  - May be based on probability, distance measurement, etc.
- Binary label
  - Strict assignment of a outlier flag.
  - Contain less information than real-value score.
  - May be based on real-values with threshold.

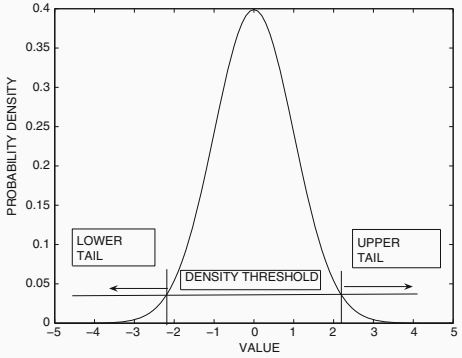
# Outlier analysis - Model types

- Extreme values
- Clustering models
- Distance-based models
- Density based models
- Probabilistic models
- Information-theoretic models

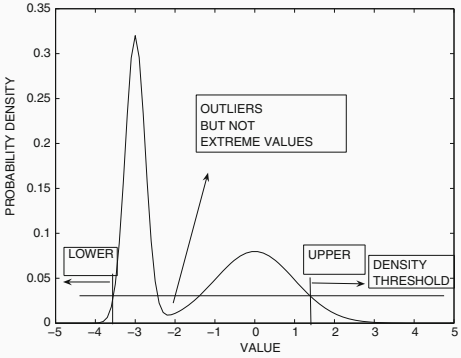
# Outlier analysis - Extreme values - Example



# Outlier analysis - Extreme values - Univariate

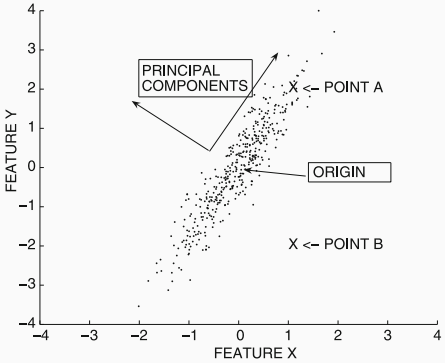



(a) Symmetric distribution

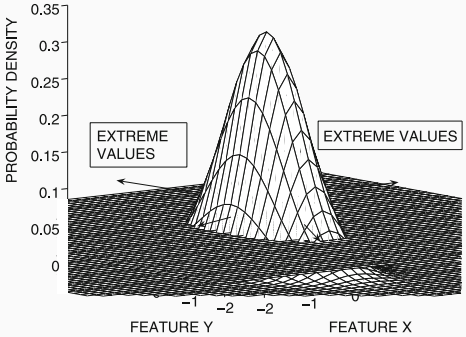



(b) Asymmetric distribution

# Outlier analysis - Extreme values - Multivariate

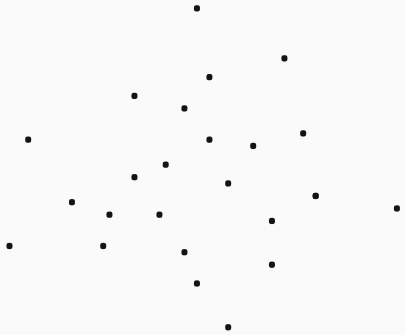



(a) Multivariate extreme values

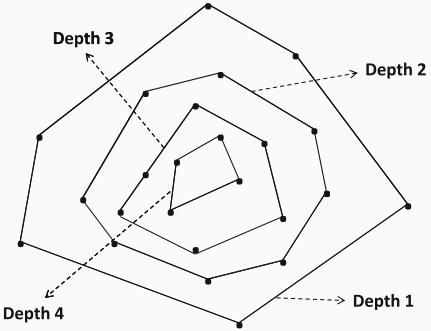



(b) Multivariate extreme values (probabilistic interpretation)

# Outlier analysis - Extreme values - Depth-based

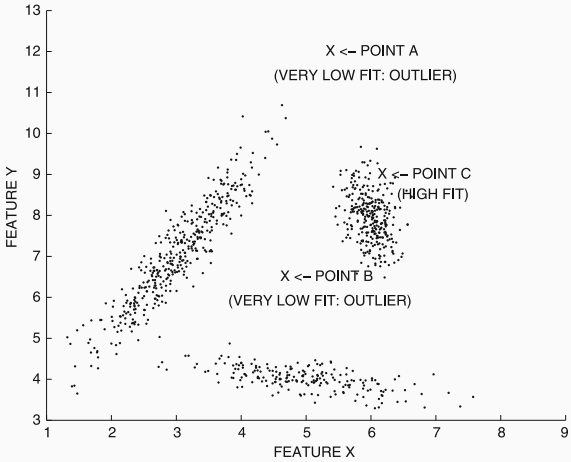


(a)




(b)

# Outlier analysis - Probability values - Likelihood

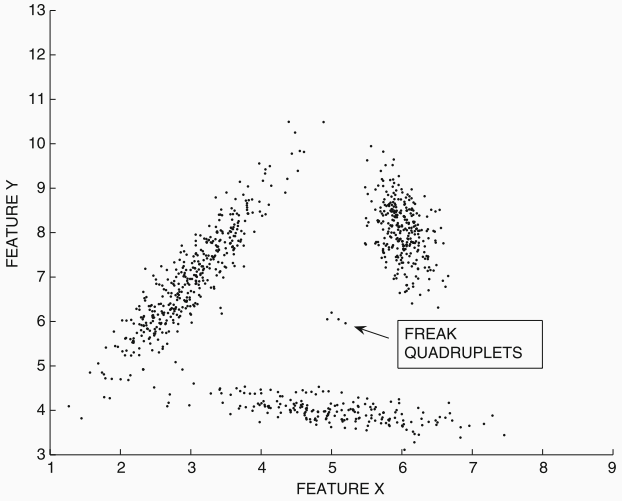


# Outlier analysis - Distance-based detection

The distance-based outlier score of an object O is its distance to its $k$-th nearest neighbor.

- The definition is based on the user defined $k$.
- The $k > 1$ helps to removes a group of outliers.
- Outlier detection methods use finer granularity than clustering methods.
- The ambient noise has lover k-nearest neighbor distance than truly isolated anomaly.
- The better granularity brings higher computational complexity.
- The speed-up techniques are used:
  - Index structures
  - Pruning tricks.

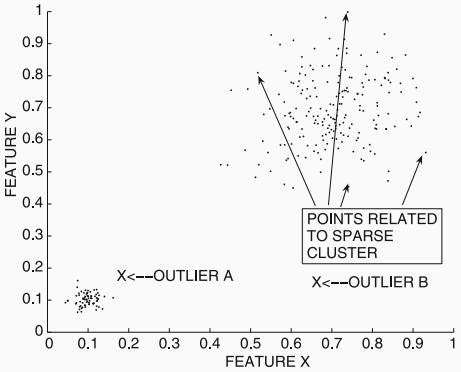
# Outlier analysis - Distance-based detection



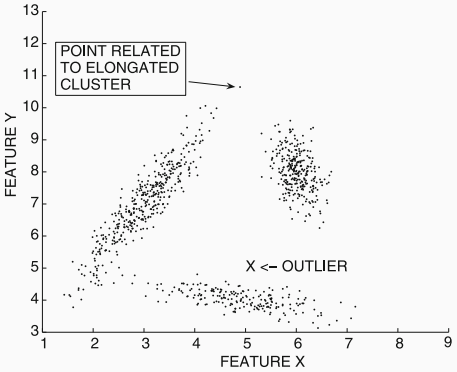
# Outlier analysis - Distance-based detection - Pruning methods

- Sampling methods
  - A sample $S$ of size $s$ is sampled from data.
  - Distances between all pairs from $D$ and $S$ are computed.
  - The complexity is $O(n \cdot s) \ll O(n^2)$.
  - The top $r$ ranked outlier in sample $S$ is determined.
  - The score of the $r$th ranked outlier is the lover bound $L$ over the the whole dataset.
  - The upper bound $V^k(X)$ for each point $X \in D - S$ is known from distances of pairs.
  - When $V^k(X)$ is not larger than the lower bound L the point X may be excluded from the testing.
  - Large number of points is removed due to this condition.
  - The remaining points $R \subseteq D - S$ is tested by outlier measure.
  - The proper ordering of $R$ and $D - S$ may significantly improve the speed of the algorithm.

# Outlier analysis - Distance-based detection - Local distance correction




(a) Varying cluster density




(b) Varying cluster shape

# Outlier analysis - Distance-based detection - Local distance correction

- Local outlier factor (LOF)
  - Adjusting the local variations in cluster density by normalization of distanced with the average point-specific distances in a data locality.
  - This approach solves the varying cluster density situation.
  - For a given point $X$ the $V^k(X)$ is its distance to its $k$-nearest neighbor.
  - The $L_k(X)$ is the set of points within the $k$-nearest neighbor distance of $X$.
  - The number of points in $L_k(X)$ is $k$ or more.
  - The reachability distance is defined as

$$R_k(X, Y) = \max\{Dist(X, Y), V^k(Y)\}$$

  - The average reachability distance of $X$ with respect to $L_k(X)$ is then

$$AR_k(X) = MEAN_{Y \in L_k(X)} R_k(X, Y)$$

  - The Local Outlier Factor $LOF_k$ is then

$$LOF_k(X) = MEAN_{Y \in L_k(X)} \frac{AR_k(X)}{AR_k(Y)}$$

# Outlier analysis - Distance-based detection - Local distance correction

- Instance-Specific Mahalanobis distance
  - The goal is to deal with the varying cluster shape.
  - A $k$-local neighborhood $L_k(X)$ with respect to the cluster shape have to be defined.
  - $L_k(X)$ is constructed with the single-linkage agglomerative approach around the point $X$.
  - A mean $\mu_k(X)$ and the covariance matrix $\Sigma_k(X)$ are computed.
  - The distance $LMaha_k(X)$ then represent the outlier score.

$$LMaha_k(X) = Maha(X, \mu_k(X), \Sigma_k(X))$$

# Outlier analysis - Density based methods

- The idea is similar to the density-based clustering.
- The main difference is that only the non-dense regions are detected.
- The points in sparse regions are reported as outliers.
- Histogram-based technique
  - Popular method for univariate data.
  - Represents the statistical distribution of points.
  - Difficult to adapt to varying density in different data locality.
  - Difficult to adapt this method to higher dimensions.
- Grid-based techniques
  - The space is partitioned into $p$ equi-width ranges.
  - The sparse regions with the density less than $\tau$ are reported as outliers.
  - It is difficult to select proper $p$.
  - The $\tau$ may be defined using univariate extreme value analysis.
  - Outlier groups may not be reported because the cluster shapes are not recognized.

# Outlier analysis - Density based methods

- Kernel-based density estimation
  - Similarly to Histogram- or Grid-based methods a local density is detected.
  - The density in each point is computed as smoothed values of a kernel functions associated with each data point.

$$f(X) = \frac{1}{n} \sum_{i=1}^{n} K_h(X - X_i)$$

  - The $h$ is a parameter of a function.
  - typical choice is the Gausian kernel with the width $h$.

$$K_h(X - X_i) = \left( \frac{1}{\sqrt{2\pi}h} \right)^d \cdot e^{-\|X - X_i\|^2 / (2h^2)}$$

Questions?