Data Mining, Lecture 6: Outlier Analysis

S. Nõmm

[1] Department of Computer Science, Tallinn University of Technology

March 7, 2016

Introduction

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.
Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature.

Applications

- Data cleaning.
- Credit card fraud.
- Network intrusion detection.

Principles

- Most outlier detection methods create a model of normal patterns.
- Outliers are defined as data points that do not naturally fit within this normal model.
- The outlierness of a data point is quantified by a numeric value, known as the outlier score.
  - Real-valued outlier score quantifies the tendency for a data point to be considered an outlier.
  - Binary label is output, indicating whether or not a data point is an outlier.

Models of the normal patterns

- **Extreme values:** A data point is an extreme value, if it lies at one of the two ends of a probability distribution. [Hawkins].
- **Clustering models:** Clustering is considered a complementary problem to outlier analysis.
- **Distance-based models:** In these cases, the k-nearest neighbor distribution of a data point is analyzed to determine whether it is an outlier. Distance-based models can be considered a more fine-grained and instance-centered version of clustering models.
- **Density-based models:** The local density of a data point is used to define its outlier score.
- **Probabilistic models:** The steps are almost analogous to those of clustering algorithms, except that the EM algorithm is used for clustering, and the probabilistic fit values are used to quantify the outlier scores of data points (instead of distance values).
- **Information-theoretic models:** Constrain the maximum deviation allowed from the normal model and then examine the difference in space requirements for constructing a model with or without a specific data point. If the difference is large, then this point is reported as an outlier.

Extreme Value analysis

- Extreme value analysis is a very specific kind of outlier analysis where the data points at the outskirts of the data are reported as outliers. Such outliers correspond to the statistical tails of probability distributions.
- All extreme values are outliers, but the reverse may not be true.
- Consider $\{1, 3, 3, 3, 50, 97, 97, 97, 100\}$.
- 1 and 100 are extreme values and therefore an outliers according to the [Hawkins]. 50 is the mean and is therefore not an extreme value but at the same time it is most isolated point and should be treated as the outlier.

Univariate Extreme Value Analysis

- Univariate extreme value analysis is related to the notion of statistical tail confidence tests. Provides a level of confidence about whether or not a specific data point is an extreme value.
- Most commonly used in the cases of normal distribution. The density function with mean μ and standard deviation σ is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Z - value.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Large positive values of z_i correspond to the upper tail, whereas large negative values correspond to the lower tail.
- The normal distribution can be directly written in terms of Z - value.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z_i^2}{2}}$$

Multivariate Extreme Values

- Defined for unimodal probability distributions with a single peak.
- Let μ be the d-dimensional mean vector of a d-dimensional data set, and Σ be its $d \times d$ covariance matrix.
- The probability density function of d - dimensional data point $\bar{X}$ is

$$f(\bar{X}) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} e^{\frac{1}{2}(\bar{X}-mu)\Sigma^{-1}(\bar{X}-mu)^T}$$

- Intuitively, this approach models the data distribution along the various uncorrelated directions as statistically independent normal distributions and standardizes them so as to provide each such direction equal importance in the outlier score.