Outline

- Basic concepts
- Statistical approaches
  - Parametric methods
  - Nonparametric methods
- Proximity-based approaches
- Reconstruction-based approaches
- Clustering and classification based approaches

General Idea

- The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers
- A parametric method assumes that the normal data objects are generated by a parametric distribution with a finite number of parameters Θ
  - The probability density function of the parametric distribution $f(x, \Theta)$ gives the probability that object x is generated by the distribution
  - The smaller this value, the more likely x is an outlier
- A nonparametric method tries to determine the model from the input data

Detection of Univariate Outliers Based on Normal Distribution

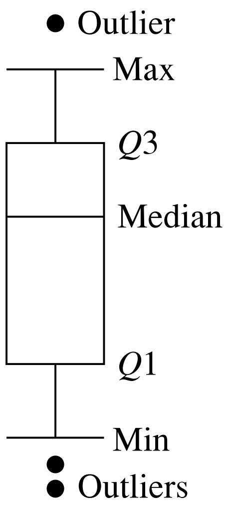
- Assumption: Data are generated from a normal distribution
- Learn the parameters of the normal (i.e., Gaussian) distribution from the input data, and identify the points with low probability as outliers
- Example: suppose a city's average temperature values in July in the last 10 years are, in value-ascending order, 24.0◦C, 28.9◦C, 28.9◦C, 29.0◦C, 29.1◦C, 29.1◦C, 29.2◦C, 29.2◦C, 29.3◦C, and 29.4◦C
- A normal distribution is determined by two parameters: the mean, μ, and the standard deviation, σ
- Use the maximum likelihood method to estimate the parameters μ and σ

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i | (\mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 28.61, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2.29$$

- 24.0◦C is an outlier, since $L(24|(28.61, 2.29)) < 0.15\%$

Boxplot Visualization



- A five-number summary
  - The smallest nonoutlier value (Min)
  - The lower quartile (Q1)
  - The median (Q2)
  - The upper quartile (Q3), and
  - The largest nonoutlier value (Max)
- The interquantile range (IQR) is defined as Q3 − Q1
- Any object that is more than 1.5 × IQR smaller than Q1 or 1.5 × IQR larger than Q3 is treated as an outlier because the region between Q1 − 1.5 × IQR and Q3 + 1.5 × IQR contains 99.3% of the objects

Multivariate Outlier Detection Using the χ^2-statistic

- The χ^2-statistic is

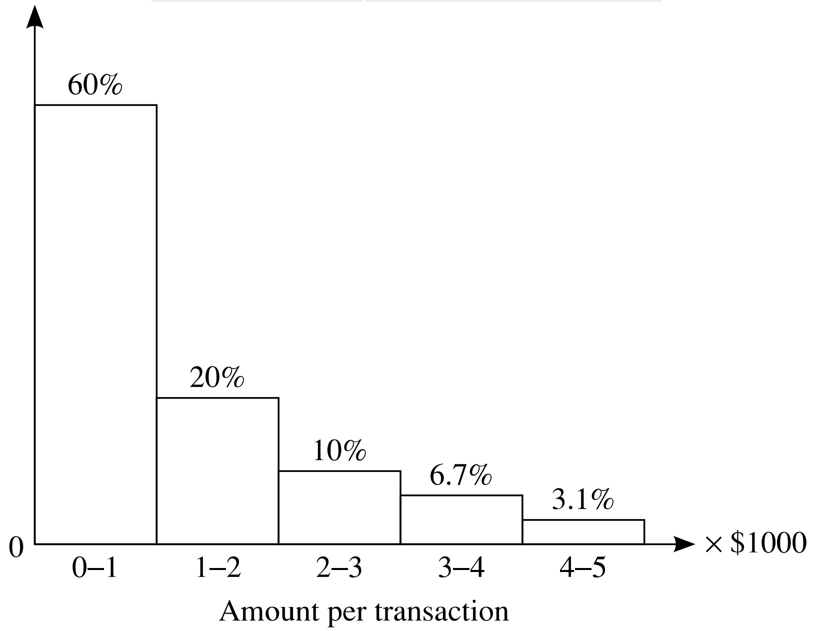
$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

  - o_i is the value of o on the i-th dimension
  - E_i is the mean of the i-dimension among all objects
- If the χ^2-statistic is large, the object is an outlier

A Nonparametric Method: Using Histogram

- Construct a histogram using the input data (training data)
- If the object falls in one of the histogram's bins, the object is regarded as normal
  - Otherwise, it is considered an outlier
- Use the histogram to assign an outlier score to an object, such as the reciprocal of the volume of the bin in which the object falls
- Drawbacks: hard to choose an appropriate bin size

Amount	Outlier score
\$7500	$\frac{1}{0.2\%} = 500$
\$385	$\frac{1}{60\%} = 1.67$



Pros and Cons of Statistical Methods

- Advantage: the outlier detection may be statistically justifiable
- Challenge: statistical methods for outlier detection on high-dimensional data
- The computational cost of statistical methods depends on the models