**Outlier detection** (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection

Outlier detection and clustering analysis are two highly related tasks. Clustering finds the majority patterns in a data set and organizes the data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns

**What Are Outliers?**

Assume that a given statistical process is used to generate a set of data objects. An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism. For ease of presentation within this chapter, we may refer to data objects that are not outliers as “normal” or expected data. Similarly, we may refer to outliers as “abnormal” data

Outliers are interesting because they are suspected of not being generated by the same mechanisms as the rest of the data. Therefore, in outlier detection, it is important to justify why the outliers detected are generated by some other mechanisms. This is often achieved by making various assumptions on the rest of the data and showing that the outliers detected violate those assumptions significantly.

In general, outliers can be classified into three categories, namely global outliers, con-textual (or conditional) outliers.

**Challenges of Outlier Detection**

**Outlier detection is useful in many applications yet faces many challenges such as the following**:

**Modeling normal objects and outliers effectively**.

Outlier detection quality highly depends on the modeling of normal (non outlier) objects and outliers. Often, building a comprehensive model for data normality is very challenging, if not impossible. This is partly because it is hard to enumerate all possible normal behaviors in an application. The border between data normality and abnormality (outliers) is often not clear cut. Instead, there can be a wide range of gray area. Consequently, while some outlier detection methods assign to each object in the input data set a label of either “normal” or “outlier,” other methods assign to each object a score measuring the “outlier-ness” of the object.

**Application-specific outlier detection**. Technically, choosing the similarity/distance measure and the relationship model to describe data objects is critical in outlier detection. Unfortunately, such choices are often application-dependent. Different applications may have very different requirements. For example, in clinic data analysis, a small deviation may be important enough to justify an outlier. In contrast, in marketing analysis, objects are often subject to larger fluctuations, and consequently a substantially larger deviation is needed to justify an outlier. Outlier detection’s high dependency on the application type makes it impossible to develop a universally applicable outlier detection method. Instead, individual outlier detection methods that are dedicated to specific applications must be developed

**Handling noise in outlier detection**.

As mentioned earlier, outliers are different from noise. It is also well known that the quality of real data sets tends to be poor. Noise often unavoidably exists in data collected in many applications. Noise may be present as deviations in attribute values or even as missing values. Low data quality and the presence of noise bring a huge challenge to outlier detection. They can distort the data, blurring the distinction between normal objects and outliers. Moreover, noise and missing data may “hide” outliers and reduce the effectiveness of outlier detection—an outlier may appear “disguised” as a noise point, and an outlier detection method may mistakenly identify a noise point as an outlier.

**Understandability.**

In some application scenarios, a user may want to not only detect outliers, but also understand why the detected objects are outliers. To meet the understandability requirement, an outlier detection method has to provide some justification of the detection. For example, a statistical method can be used to justify the degree to which an object may be an outlier based on the likelihood that the object was generated by the same mechanism that generated the majority of the data. The smaller the likelihood, the more unlikely the object was generated by the same mechanism, and the more likely the object is an outlier

**Outlier Detection Methods**

**Supervised, Semi-Supervised, and Unsupervised Methods**

**Supervised Methods**

Supervised methods model data normality and abnormality. Domain experts examine and label a sample of the underlying data. Outlier detection can then be modeled as a classification problem (Chapters 8 and 9). The task is to learn a classifier that can recognize outliers. The sample is used for training and testing. In some applications, the experts may label just the normal objects, and any other objects not matching the model of normal objects are reported as outliers. Other methods model the outliers and treat objects not matching the model of outliers as normal

Although many classification methods can be applied, challenges to supervised outlier detection include the following:

In many outlier detection applications, catching as many outliers as possible (i.e., the sensitivity or recall of outlier detection) is far more important than not mislabeling normal objects as outliers. Consequently, when a classification method is used for supervised outlier detection, it has to be interpreted appropriately so as to consider the application interest on recall.

**Unsupervised Methods**

Unsupervised outlier detection methods make an implicit assumption: The normal objects are somewhat “clustered.” In other words, an unsupervised outlier detection method expects that normal objects follow a pattern far more frequently than outliers. Normal objects do not have to fall into one group sharing high similarity. Instead, they can form multiple groups, where each group has distinct features. However, an outlier is expected to occur far away in feature space from any of those groups of normal objects.

In some applications, normal objects are diversely distributed, and many such objects do not follow strong patterns. For instance, in some intrusion detection and computer virus detection problems, normal activities are very diverse and many do not fall into high-quality clusters. In such scenarios, unsupervised methods may have a high false positive rate—they may mislabel many normal objects as outliers (intrusions or viruses in these applications), and let many actual outliers go undetected. Due to the high similarity between intrusions and viruses (i.e., they have to attack key resources in the target systems), modeling outliers using supervised methods may be far more effective

**Semi-Supervised Methods**

We may encounter cases where only a small set of the normal and/or outlier objects are labeled, but most of the data are unlabeled. Semi-supervised outlier detection methods were developed to tackle such scenarios

For example, when some labeled normal objects are available, we can use them, together with unlabeled objects that are close by, to train a model for normal objects. The model of normal objects then can be used to detect outliers—those objects not fitting the model of normal objects are classified as outliers. If only some labeled outliers are available, semi-supervised outlier detection is trickier. A small number of labeled outliers are unlikely to represent all the possible outliers. Therefore, building a model for outliers based on only a few labeled outliers is unlikely to be effective. To improve the quality of outlier detection, we can get help from models for normal objects learned from unsupervised methods.

**Statistical Methods Statistical methods** (also known as model-based methods) make assumptions of data normality. They assume that normal data objects are generated by a statistical (stochastic) model, and that data not following the model are outliers

**Proximity-based methods** assume that an object is an outlier if the nearest neighbors of the object are far away in feature space, that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set.

**Clustering-based methods** assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters

**Statistical Approaches**

Statistical methods for outlier detection can be divided into two major categories: parametric methods and nonparametric methods, according to how the models are specified and learned

**A parametric method** assumes that the normal data objects are generated by a parametric distribution with parameter 2. The probability density function of the parametric distribution f (x,2) 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** does not assume an a priori statistical model. Instead, a nonparametric method tries to determine the model from the input data. Note that most nonparametric methods do not assume that the model is completely parameterfree.

**Parametric Methods**

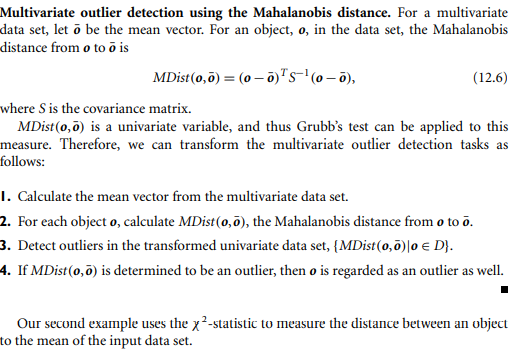
**Detection of Univariate Outliers Based on Normal Distribution**

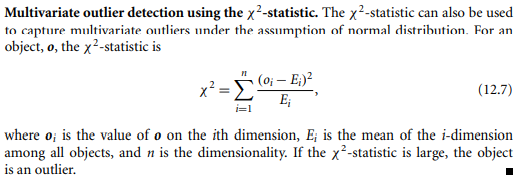
Data involving only one attribute or variable are called univariate data. For simplicity, we often choose to assume that data are generated from a normal distribution. We can then learn the parameters of the normal distribution from the input data, and identify the points with low probability as outliers

**Univariate outlier detection using maximum likelihood.** 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. Let’s assume that the average temperature follows a normal distribution, which is determined by two parameters: the mean, µ, and the standard deviation, σ

**Detection of Multivariate Outliers**

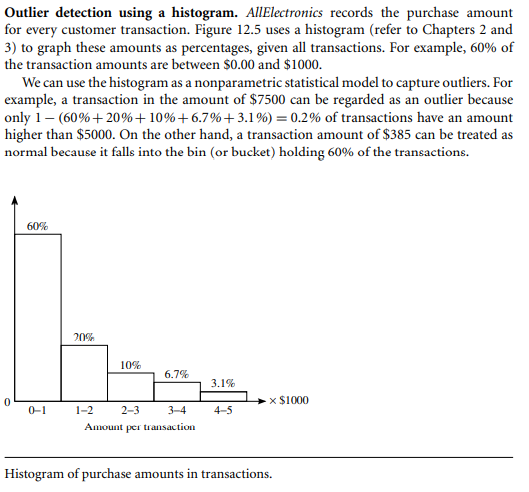
Data involving two or more attributes or variables are multivariate data. Many univariate outlier detection methods can be extended to handle multivariate data. The central idea is to transform the multivariate outlier detection task into a univariate outlier detection problem.

****

****

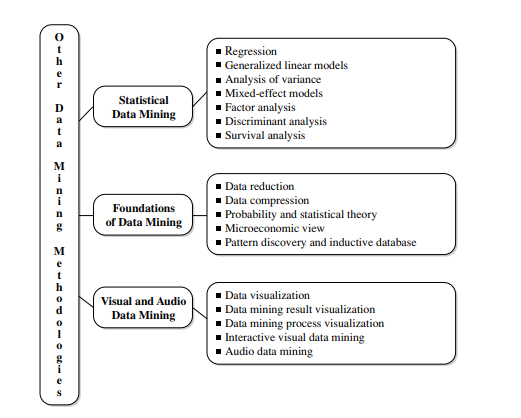
**Nonparametric Methods**

In nonparametric methods for outlier detection, the model of “normal data” is learned from the input data, rather than assuming one a priori. Nonparametric methods often make fewer assumptions about the data, and thus can be applicable in more scenarios

****

**Statistical Data Mining**

The data mining techniques described in this book are primarily drawn from computer science disciplines, including data mining, machine learning, data warehousing, and algorithms. They are designed for the efficient handling of huge amounts of data that are typically multidimensional and possibly of various complex types. There are, however, many well-established statistical techniques for data analysis, particularly for numeric data. These techniques have been applied extensively to scientific data (e.g., data from experiments in physics, engineering, manufacturing, psychology, and medicine), as well as to data from economics and the social sciences.

****

A thorough discussion of major statistical methods for data analysis is beyond the scope of this book; however, several methods are mentioned here for the sake of completeness. Pointers to these techniques are provided in the bibliographic notes (Section 13.8).

**Regression: In** general, these methods are used to predict the value of a response (dependent) variable from one or more predictor (independent) variables, where the variables are numeric. There are various forms of regression, such as linear, multiple, weighted, polynomial, nonparametric, and robust (robust methods are useful when errors fail to satisfy normalcy conditions or when the data contain significant outliers).

**Generalized linear models**: These models, and their generalization (generalized additive models), allow a categorical (nominal) response variable (or some transformation

**Analysis of variance**: These techniques analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors). In general, an ANOVA (single-factor analysis of variance) problem involves a comparison of k population or treatment means to determine if at least two of the means are different. More complex ANOVA problems also exist.

**Mixed-effect models**: These models are for analyzing grouped data—data that can be classified according to one or more grouping variables. They typically describe relationships between a response variable and some covariates in data grouped according to one or more factors. Common areas of application include multilevel data, repeated measures data, block designs, and longitudinal data.

**Discriminant analysis**: This technique is used to predict a categorical response variable. Unlike generalized linear models, it assumes that the independent variables follow a multivariate normal distribution. The procedure attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable. Discriminant analysis is commonly used in social science

**Data Mining for Financial Data Analysis- 607**

**Data Mining for Intrusion Detection and Prevention- 614**

**Data Mining and Recommender Systems- 615**