# 5

# STATISTICAL METHODS

**Chapter Objectives**

- Explain methods of statistical inference commonly used in data-mining applications.
- Identify different statistical parameters for assessing differences in data sets.
- Describe the components and the basic principles of Naïve Bayesian classifier and the logistic regression method.
- Introduce log-linear models using correspondence analysis of contingency tables.
- Discuss the concepts of analysis of variance (ANOVA) and linear discriminant analysis (LDA) of multidimensional samples.

Statistics is the science of collecting and organizing data and drawing conclusions from data sets. The organization and description of the general characteristics of data sets is the subject area of *descriptive statistics*. How to draw conclusions from data is the subject of *statistical inference*. In this chapter, the emphasis is on the basic principles

*Data Mining: Concepts, Models, Methods, and Algorithms*, Second Edition. Mehmed Kantardzic.

of statistical inference; other related topics will be described briefly enough to understand the basic concepts.

Statistical data analysis is the most well-established set of methodologies for data mining. Historically, the first computer-based applications of data analysis were developed with the support of statisticians. Ranging from one-dimensional data analysis to multivariate data analysis, statistics offered a variety of methods for data mining, including different types of regression and discriminant analysis. In this short overview of statistical methods that support the data-mining process, we will not cover all approaches and methodologies; a selection has been made of the techniques used most often in real-world data-mining applications.

## 5.1 STATISTICAL INFERENCE

The totality of the observations with which we are concerned in statistical analysis, whether their number is finite or infinite, constitutes what we call a *population*. The term refers to anything of statistical interest, whether it is a group of people, objects, or events. The number of observations in the population is defined as the size of the population. In general, populations may be finite or infinite, but some finite populations are so large that, in theory, we assume them to be infinite.

In the field of statistical inference, we are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average length of the life of a certain brand of light bulbs, it would be practically impossible to test all such bulbs. Therefore, we must depend on a subset of observations from the population for most statistical-analysis applications. In statistics, a subset of a population is called a *sample* and it describes a finite data set of n-dimensional vectors. Throughout this book, we will simply call this subset of population *data set*, to eliminate confusion between the two definitions of sample: one (explained earlier) denoting the description of a single entity in the population, and the other (given here) referring to the subset of a population. From a given data set, we build a statistical model of the population that will help us to make inferences concerning that same population. If our inferences from the data set are to be valid, we must obtain samples that are representative of the population. Very often, we are tempted to choose a data set by selecting the most convenient members of the population. But such an approach may lead to erroneous inferences concerning the population. Any sampling procedure that produces inferences that consistently overestimate or underestimate some characteristics of the population is said to be biased. To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a random data set in the sense that the observations are made independently and at random. The main purpose of selecting random samples is to elicit information about unknown population parameters.

The relation between data sets and the system they describe may be used for inductive reasoning: from observed data to knowledge of a (partially) unknown system. Statistical inference is the main form of reasoning relevant to data analysis. The theory of statistical inference consists of those methods by which one makes inferences or

generalizations about a population. These methods may be categorized into two major areas: *estimation* and *tests of hypotheses*.

In *estimation,* one wants to come up with a plausible value or a range of plausible values for the unknown parameters of the system. The goal is to gain information from a data set T in order to estimate one or more parameters w belonging to the model of the real-world system f(X, w). A data set T is described by the ordered n-tuples of values for variables: $X = \{X_1, X_2, \ldots, X_n\}$ (attributes of entities in population):

$$T = \{(x_{11}, \ldots, x_{1n}), (x_{21}, \ldots, x_{2n}), \ldots, (x_{m1}, \ldots, x_{mn})\}$$

It can be organized in a tabular form as a set of samples with its corresponding feature values. Once the parameters of the model are estimated, we can use them to make predictions about the random variable Y from the initial set of attributes $Y \in X$, based on other variables or sets of variables $X^* = X - Y$. If Y is numeric, we speak about *regression*, and if it takes its values from a discrete, unordered data set, we speak about *classification*.

Once we have obtained estimates for the model parameters w from some data set T, we may use the resulting model (analytically given as a function f[X*, w]) to make predictions about Y when we know the corresponding value of the vector X*. The difference between the prediction f(X*, w) and the real value Y is called the prediction error. It should preferably take values close to 0. A natural quality measure of a model f(X*, w), as a predictor of Y, is the expected mean-squared error for the entire data set T:

$$E_T[(Y - f[X^*, w])^2].$$

In *statistical testing*, on the other hand, one has to decide whether a hypothesis concerning the value of the population characteristic should be accepted or rejected in light of an analysis of the data set. A statistical hypothesis is an assertion or conjecture concerning one or more populations. The truth or falsity of a statistical hypothesis can never be known with absolute certainty, unless we examine the entire population. This, of course, would be impractical in most situations, sometimes even impossible. Instead, we test a hypothesis on a randomly selected data set. Evidence from the data set that is inconsistent with the stated hypothesis leads to a rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance, or more precisely, it implies that the data do not contain sufficient evidence to refute it. The structure of hypothesis testing is formulated with the use of the term *null hypothesis*. This refers to any hypothesis that we wish to test and is denoted by $H_0$. $H_0$ is only rejected if the given data set, on the basis of the applied statistical tests, contains strong evidence that the hypothesis is not true. The rejection of $H_0$ leads to the acceptance of an alternative hypothesis about the population.

In this chapter, some statistical estimation and hypothesis-testing methods are described in great detail. These methods have been selected primarily based on the applicability of the technique in a data-mining process on a large data set.

## 5.2 ASSESSING DIFFERENCES IN DATA SETS

For many data-mining tasks, it would be useful to learn the more general characteristics about the given data set, regarding both central tendency and data dispersion. These simple parameters of data sets are obvious descriptors for assessing differences between different data sets. Typical measures of central tendency include *mean*, *median*, and *mode*, while measures of data dispersion include *variance* and *standard deviation*.

The most common and effective numeric measure of the center of the data set is the *mean* value (also called the arithmetic mean). For the set of n numeric values $x_1$, $x_2$, . . . , $x_n$, for the given feature X, the mean is

$$mean = 1/n \sum_{i=1}^{n} x_i$$

and it is a built-in function (like all other descriptive statistical measures) in most modern, statistical software tools. For each numeric feature in the n-dimensional set of samples, it is possible to calculate the mean value as a central tendency characteristic for this feature. Sometimes, each value $x_i$ in a set may be associated with a weight $w_i$, which reflects the frequency of occurrence, significance, or importance attached to the value. In this case, the weighted arithmetic mean or the weighted average value is

$$mean = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$$

Although the mean is the most useful quantity that we use to describe a set of data, it is not the only one. For skewed data sets, a better measure of the center of data is the *median*. It is the middle value of the ordered set of feature values if the set consists of an odd number of elements and it is the average of the middle two values if the number of elements in the set is even. If $x_1$, $x_2$, . . . , $x_n$ represents a data set of size n, arranged in increasing order of magnitude, then the median is defined by

$$median = \begin{cases} x_{(n+1)/2} & \text{if n is odd} \\ (x_{n/2} + x_{(n/2)+1})/2 & \text{if n is even} \end{cases}$$

Another measure of the central tendency of a data set is the *mode*. The mode for the set of data is the value that occurs most frequently in the set. While mean and median are characteristics of primarily numeric data sets, the mode may be applied also to categorical data, but it has to be interpreted carefully because the data are not ordered. It is possible for the greatest frequency to correspond to several different values in a data set. That results in more than one mode for a given data set. Therefore, we classify data sets as unimodal (with only one mode) and multimodal (with two or more modes). Multimodal data sets may be precisely defined as bimodal, trimodal, and so on. For unimodal frequency curves that are moderately asymmetrical, we have the following useful empirical relation for numeric data sets

$$\text{mean} - \text{mode} \leq 3 \times (\text{mean} - \text{median})$$

that may be used for an analysis of data set distribution and the estimation of one central-tendency measure based on the other two.

As an example, let us analyze these three measures on the simple data set T that has the following numeric values:

$$T = \{3, 5, 2, 9, 0, 7, 3, 6\}.$$

After a sorting process the same data set is given as

$$T = \{0, 2, 3, 3, 5, 6, 7, 9\}.$$

The corresponding descriptive statistical measures for central tendency are

$$mean_T = (0+2+3+3+5+6+7+9)/8 = 4.375$$
$$median_T = (3+5)/2 = 4$$
$$mode_T = 3$$

The degree to which numeric data tend to spread is called dispersion of the data, and the most common measures of dispersion are the *standard deviation* $\sigma$ and the *variance* $\sigma^2$. The variance of n numeric values $x_1, x_2, \ldots, x_n$ is

$$\sigma^2 = (1/(n-1)) \sum_{i=1}^{n} (x_i - mean)^2$$

The standard deviation $\sigma$ is the square root of the variance $\sigma^2$. The basic properties of the standard deviation $\sigma$ as a measure of spread are

1. $\sigma$ measures spread about the *mean* and should be used only when the *mean* is chosen as a measure of the center.
2. $\sigma = 0$ only when there is no spread in the data, that is, when all measurements have the same value. Otherwise $\sigma > 0$.

For the data set given in our example, *variance* $\sigma^2$ and *standard deviation* $\sigma$ are

$$\sigma^2 = 1/7 \sum_{i=1}^{8} (x_i - 4.375)^2$$
$$\sigma^2 = 8.5532$$
$$\sigma = 2.9246$$

In many statistical software tools, a popularly used visualization tool of descriptive statistical measures for central tendency and dispersion is a *boxplot* that is typically determined by the mean value, variance, and sometimes max and min values of the data set. In our example, the minimal and maximal values in the T set are $\min_T = 0$,

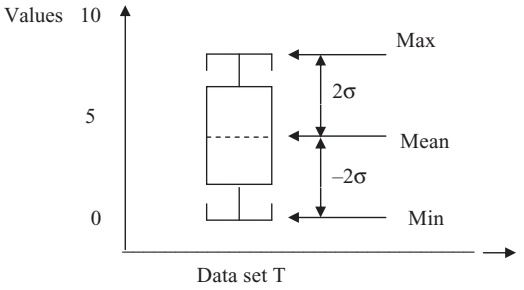


Figure 5.1. A boxplot representation of the data set T based on mean value, variance, and min and max values.

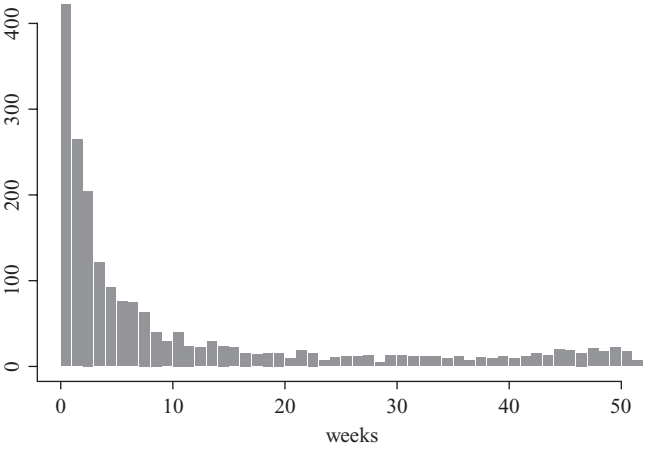


Figure 5.2. Displaying a single-feature distribution.

and $\max_T = 9$. Graphical representation of statistical descriptors for the data set T has a form of a boxplot, given in Figure 5.1.

Analyzing large data sets requires proper understanding of the data in advance. This would help domain experts to influence the data-mining process and to properly evaluate the results of a data-mining application. Central tendency measures for a data set are valuable only for some specific distributions of data values. Therefore, it is important to know the characteristics of a distribution for a data set we are analyzing. The distribution of values in a data set is described according to the spread of its values. Usually, this is best done using a histogram representation; an example is given in Figure 5.2. In addition to quantifying the distribution of values for each feature, it is also important to know the global character of the distributions and all specifics. Knowing that data set has a classic bell curve empowers researchers to use a broad range of traditional statistical techniques for data assessment. But in many practical cases, the distributions are skewed or multimodal, and traditional interpretations of concepts such as mean value or standard deviation do not make sense.

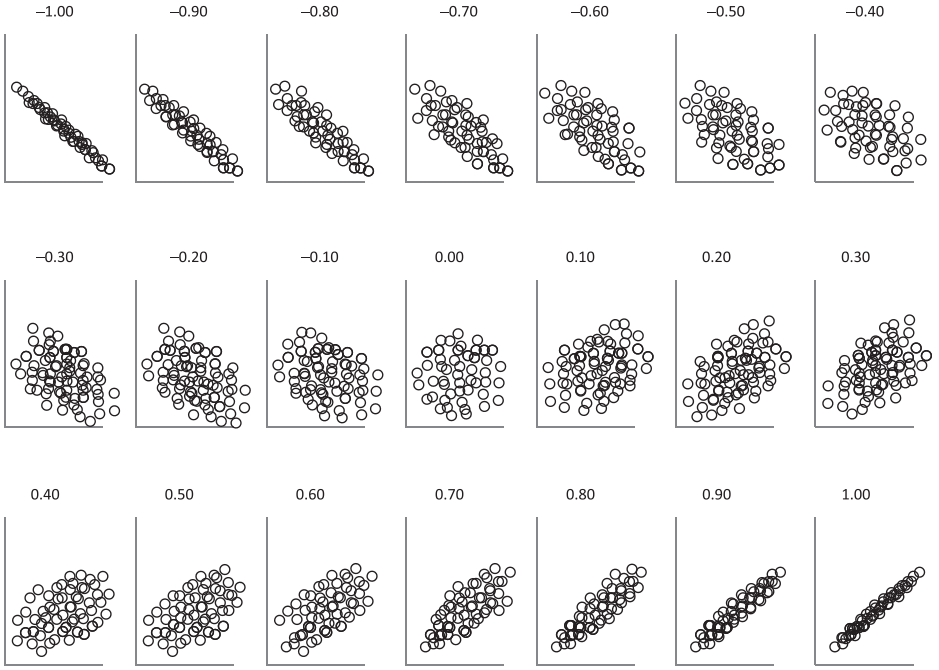


Figure 5.3. Scatter plots showing the correlation between features from –1 to 1.

Part of the assessment process is determining relations between features in a data set. Simple visualization through the scatter plots gives initial estimation of these relations. Figure 5.3 shows part of the integrated scatter plot where each pair of features is compared. This visualization technique is available in most integrated data-mining tools. Quantification of these relations is given through the correlation factor.

These visualization processes are part of the data-understanding phase, and they are essential in preparing better for data mining. This human interpretation helps to obtain a general view of the data. It is also possible to identify abnormal or interesting characteristics, such as anomalies.

## 5.3 BAYESIAN INFERENCE

It is not hard to imagine situations in which the data are not the only available source of information about the population or about the system to be modeled. The Bayesian method provides a principled way to incorporate this external information into the data-analysis process. This process starts with an already given probability distribution for the analyzed data set. As this distribution is given before any data is considered, it is called a *prior distribution*. The new data set updates this prior distribution into a *posterior distribution*. The basic tool for this updating is the Bayes theorem.

The Bayes theorem represents a theoretical background for a statistical approach to inductive-inferencing classification problems. We will explain first the basic concepts

defined in the Bayes theorem, and then, use this theorem in the explanation of the Naïve Bayesian classification process, or the *simple Bayesian classifier*.

Let X be a data sample whose class label is unknown. Let H be some hypothesis: such that the data sample X belongs to a specific class C. We want to determine P(H/X), the probability that the hypothesis H holds given the observed data sample X. P(H/X) is the posterior probability representing our confidence in the hypothesis after X is given. In contrast, P(H) is the prior probability of H for any sample, regardless of how the data in the sample look. The posterior probability P(H/X) is based on more information then the prior probability P(H). The Bayesian theorem provides a way of calculating the posterior probability P(H/X) using probabilities P(H), P(X), and P(X/H). The basic relation is

$$P(H/X) = [P(X/H) \cdot P(H)]/P(X)$$

Suppose now that there is a set of m samples $S = \{S_1, S_2, \ldots, S_m\}$ (the training data set) where every sample $S_i$ is represented as an n-dimensional vector $\{x_1, x_2, \ldots, x_n\}$. Values $x_i$ correspond to attributes $A_1, A_2, \ldots, A_n$, respectively. Also, there are k classes $C_1, C_2, \ldots, C_k$, and every sample belongs to one of these classes. Given an additional data sample X (its class is unknown), it is possible to predict the class for X using the highest conditional probability $P(C_i/X)$, where $i = 1, \ldots, k$. That is the basic idea of Naïve Bayesian classifier. These probabilities are computed using Bayes theorem:

$$P(C_i/X) = [P(X/C_i) \cdot P(C_i)]/P(X)$$

As P(X) is constant for all classes, only the product $P(X/C_i) \cdot P(C_i)$ needs to be maximized. We compute the prior probabilities of the class as

$$P(C_i) = \text{number of training samples of class } C_i / m$$

where m is total number of training samples.

Because the computation of $P(X/C_i)$ is extremely complex, especially for large data sets, the naïve assumption of conditional independence between attributes is made. Using this assumption, we can express $P(X/C_i)$ as a product:

$$P(X/C_i) = \prod_{t=1}^{n} P(x_t/C_i)$$

where $x_t$ are values for attributes in the sample X. The probabilities $P(x_t/C_i)$ can be estimated from the training data set.

A simple example will show that the Naïve Bayesian classification is a computationally simple process even for large training data sets. Given a training data set of seven four-dimensional samples (Table 5.1), it is necessary to predict classification of the new sample $X = \{1, 2, 2, \text{class} = ?\}$. For each sample, $A_1$, $A_2$, and $A_3$ are input dimensions and C is the output classification.

In our example, we need to maximize the product $P(X/C_i) \cdot P(C_i)$ for $i = 1,2$ because there are only two classes. First, we compute prior probabilities $P(C_i)$ of the class:

TABLE 5.1. Training Data Set for a Classification Using Naïve Bayesian Classifier

| Sample | Attribute1 | Attribute2 | Attribute3 | Class |
|---|---|---|---|---|
| | $A_1$ | $A_2$ | $A_3$ | C |
| 1 | 1 | 2 | 1 | 1 |
| 2 | 0 | 0 | 1 | 1 |
| 3 | 2 | 1 | 2 | 2 |
| 4 | 1 | 2 | 1 | 2 |
| 5 | 0 | 1 | 2 | 1 |
| 6 | 2 | 2 | 2 | 2 |
| 7 | 1 | 0 | 1 | 1 |

$$P(C = 1) = 4/7 = 0.5714$$

$$P(C = 2) = 3/7 = 0.4286$$

Second, we compute conditional probabilities $P(x_i/C_i)$ for every attribute value given in the new sample X = {1, 2, 2, C = ?}, (or more precisely, X = {$A_1$ = 1, $A_2$ = 2, $A_3$ = 2, C = ?}) using training data sets:

$$P(A_1 = 1/C = 1) = 2/4 = 0.50$$

$$P(A_1 = 1/C = 2) = 1/3 = 0.33$$

$$P(A_2 = 2/C = 1) = 1/4 = 0.25$$

$$P(A_2 = 2/C = 2) = 2/3 = 0.66$$

$$P(A_3 = 2/C = 1) = 1/4 = 0.25$$

$$P(A_3 = 2/C = 2) = 2/3 = 0.66$$

Under the assumption of conditional independence of attributes, the conditional probabilities $P(X/C_i)$ will be

$$P(X/C = 1) = P(A_1 = 1/C = 1) \cdot P(A_2 = 2/C = 1) \cdot P(A_3 = 2/C = 1) = \\ = 0.50 \cdot 0.25 \cdot 0.25 = 0.03125$$

$$P(X/C = 2) = P(A_1 = 1/C = 2) \cdot P(A_2 = 2/C = 2) \cdot P(A_3 = 2/C = 2) = \\ = 0.33 \cdot 0.66 \cdot 0.66 = 0.14375$$

Finally, multiplying these conditional probabilities with corresponding a priori probabilities, we can obtain values proportional (≈) to $P(C_i/X)$ and find their maximum:

$$P(C_1/X) \approx P(X/C = 1) \cdot P(C = 1) = 0.03125 \cdot 0.5714 = 0.0179$$

$$P(C_2 / X) \approx P(X / C = 2) \cdot P(C = 2) = 0.14375 \cdot 0.4286 = 0.0616$$
$$\Downarrow$$
$$P(C_2 / X) = \text{Max}\{P(C_1 / X), P(C_2 / X)\} = \{0.0179, 0.0616\}$$

Based on the previous two values that are the final results of the Naive Bayesian classifier, we can predict that the new sample X belongs to the class C = 2. The product of probabilities for this class P(X/C = 2) · P(C = 2) is higher, and therefore P(C = 2/X) is higher because it is directly proportional to the computed probability product.

In theory, the Bayesian classifier has the minimum error rate compared with all other classifiers developed in data mining. In practice, however, this is not always the case because of inaccuracies in the assumptions of attributes and class-conditional independence.

## 5.4 PREDICTIVE REGRESSION

The prediction of continuous values can be modeled by a statistical technique called *regression*. The objective of regression analysis is to determine the best model that can relate the output variable to various input variables. More formally, regression analysis is the process of determining how a variable Y is related to one or more other variables $x_1, x_2, \ldots, x_n$. Y is usually called the response output, or dependent variable, and $x_i$-s are inputs, regressors, explanatory variables, or independent variables. Common reasons for performing regression analysis include

1. the output is expensive to measure but the inputs are not, and so a cheap prediction of the output is sought;
2. the values of the inputs are known before the output is known, and a working prediction of the output is required;
3. controlling the input values, we can predict the behavior of corresponding outputs; and
4. there might be a causal link between some of the inputs and the output, and we want to identify the links.

Before explaining regression technique in details, let us explain the main differences between two concepts: interpolation and regression. In both cases training data set $X = \{x^t, r^t\}_{t=1,N}$ is given where $x^t$ are input features and output value $r^t \in R$.

- If there is *no noise* in the data set, the task is *interpolation*. We would like to find a function f(x) that passes through all these training points such that we have $r^t = f(x^t)$. In polynomial interpolation, given N points, we found that we can use (N − 1) degree polynomial to predict exact output r for any input x.
- In *regression*, *there is noise* ε added to the output of the unknown function f: $r^t = f(x^t) + \varepsilon$. The explanation for noise is that there are extra hidden variables $z^t$ that we cannot observe. We would like to approximate the output $r^t = f(x^t, z^t)$ by

our model $g(x^t)$, not only for present training data but for data in future. We are minimizing empirical error: $E(g/x) = 1/N \Sigma (r^t - g[x^t])^2$ for t = 1 to N.

Generalized linear regression models are currently the most frequently applied statistical techniques. They are used to describe the relationship between the trend of one variable and the values taken by several other variables. Modeling this type of relationship is often called linear regression. Fitting models is not the only task in statistical modeling. We often want to select one of several possible models as being the most appropriate. An objective method for choosing between different models is called ANOVA, and it is described in Section 5.5.

The relationship that fits a set of data is characterized by a prediction model called a *regression equation*. The most widely used form of the regression model is the general linear model formally written as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \ldots + \beta_n \cdot X_n$$

Applying this equation to each of the given samples we obtain a new set of equations

$$y_j = \alpha + \beta_1 \cdot x_{1j} + \beta_2 \cdot x_{2j} + \beta_3 \cdot x_{3j} + \ldots + \beta_n \cdot x_{nj} + \varepsilon_j \qquad j = 1, \ldots, m$$

where $\varepsilon_j$'s are errors of regression for each of m given samples. The linear model is called linear because the expected value of $y_j$ is a linear function: the weighted sum of input values.

Linear regression with one input variable is the simplest form of regression. It models a random variable Y (called a response variable) as a linear function of another random variable X (called a predictor variable). Given n samples or data points of the form $(x_1, y_1)$, $(x_2, y_2)$, . . . , $(x_n, y_n)$, where $x_i \in X$ and $y_i \in Y$, linear regression can be expressed as

$$Y = \alpha + \beta \cdot X$$

where $\alpha$ and $\beta$ are regression coefficients. With the assumption that the variance of Y is a constant, these coefficients can be solved by the method of least squares, which minimizes the error between the actual data points and the estimated line. The residual sum of squares is often called the sum of squares of the errors about the regression line and it is denoted by SSE (sum of squares error):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i\,')^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

where $y_i$ is the real output value given in the data set, and $y_i$' is a response value obtained from the model. Differentiating SSE with respect to $\alpha$ and $\beta$, we have

$$\partial(SEE)/\partial\alpha = -2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)$$

$$\partial(SEE)/\partial\beta = -2\sum_{i=1}^{n} ((y_i - \alpha - \beta x_i) \cdot x_i)$$

TABLE 5.2. A Database for the Application of Regression Methods

| A | B |
|---|---|
| 1 | 3 |
| 8 | 9 |
| 11 | 11 |
| 4 | 5 |
| 3 | 2 |

Setting the partial derivatives equal to 0 (minimization of the total error) and rearranging the terms, we obtain the equations

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i$$

which may be solved simultaneously to yield the computing formulas for $\alpha$ and $\beta$. Using standard relations for the mean values, regression coefficients for this simple case of optimization are

$$\beta = \left[ \sum_{i=1}^{n} (x_i - \text{mean}_x) \cdot (y_i - \text{mean}_y) \right] / \left[ \sum_{i=1}^{n} (x_i - \text{mean}_x)^2 \right]$$

$$\alpha = \text{mean}_y - \beta \cdot \text{mean}_x$$

where $\text{mean}_x$ and $\text{mean}_y$ are the mean values for random variables X and Y given in a training data set. It is important to remember that our values of $\alpha$ and $\beta$, based on a given data set, are only estimates of the true parameters for the entire population. The equation $y = \alpha + \beta x$ may be used to predict the mean response $y_0$ for the given input $x_0$, which is not necessarily from the initial set of samples.

For example, if the sample data set is given in the form of a table (Table 5.2), and we are analyzing the linear regression between two variables (predictor variable A and response variable B), then the linear regression can be expressed as

$$B = \alpha + \beta \cdot A$$

where $\alpha$ and $\beta$ coefficients can be calculated based on previous formulas (using $\text{mean}_A = 5.4$, and $\text{mean}_B = 6$), and they have the values

$$\alpha = 0.8$$

$$\beta = 0.92$$

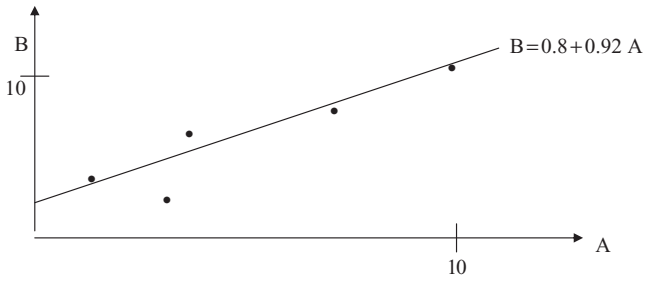


Figure 5.4. Linear regression for the data set given in Table 5.2.

The optimal regression line is

$$B = 0.8 + 0.92 \cdot A$$

The initial data set and the regression line are graphically represented in Figure 5.4 as a set of points and a corresponding line.

*Multiple regression* is an extension of linear regression, and involves more than one predictor variable. The response variable Y is modeled as a linear function of several predictor variables. For example, if the predictor attributes are $X_1$, $X_2$, and $X_3$, then the multiple linear regression is expressed as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

where $\alpha$, $\beta_1$, $\beta_2$, and $\beta_3$ are coefficients that are found by using the method of least squares. For a linear regression model with more than two input variables, it is useful to analyze the process of determining $\beta$ parameters through a matrix calculation:

$$Y = \beta \cdot X$$

where $\beta = \{\beta_0, \beta_1, \ldots, \beta_n\}$, $\beta_0 = \alpha$, and X and Y are input and output matrices for a given training data set. The residual sum of the squares of errors SSE will also have the matrix representation

$$SSE = (Y - \beta \cdot X) \cdot (Y - \beta \cdot X)$$

and after optimization

$$\partial(SSE)/\partial\beta = 0 \Rightarrow (X' \cdot X) \cdot \beta = X' \cdot Y$$

the final $\beta$ vector satisfies the matrix equation

$$\beta = (X' \cdot X)^{-1}(X' \cdot Y)$$

where $\beta$ is the vector of estimated coefficients in a linear regression. Matrices X and Y have the same dimensions as the training data set. Therefore, an optimal solution for $\beta$ vector is relatively easy to find in problems with several hundreds of training samples.

TABLE 5.3. Some Useful Transformations to Linearize Regression

| Function | Proper Transformation | Form of Simple Linear Regression |
|---|---|---|
| Exponential: $Y = \alpha e^{\beta x}$ | $Y^* = \ln Y$ | Regress Y* against x |
| Power: $Y = \alpha x^{\beta}$ | $Y^* = \log Y$; $x^* = \log x$ | Regress Y* against x* |
| Reciprocal: $Y = \alpha + \beta(1/x)$ | $x^* = 1/x$ | Regress Y against x* |
| Hyperbolic: $Y = x/(\alpha + \beta x)$ | $Y^* = 1/Y$; $x^* = 1/x$ | Regress Y* against x* |

For real-world data-mining problems, the number of samples may increase to several millions. In these situations, because of the extreme dimensions of matrices and the exponentially increased complexity of the algorithm, it is necessary to find modifications and/or approximations in the algorithm, or to use totally different regression methods.

There is a large class of regression problems, initially nonlinear, that can be converted into the form of the general linear model. For example, a polynomial relationship such as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 X_3 + \beta_4 \cdot X_2 X_3$$

can be converted to the linear form by setting new variables $X_4 = X_1 \cdot X_3$ and $X_5 = X_2 \cdot X_3$. Also, polynomial regression can be modeled by adding polynomial terms to the basic linear model. For example, a cubic polynomial curve has a form

$$Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot X_2 + \beta_3 \cdot X^3$$

By applying transformation to the predictor variables ($X_1 = X$, $X_2 = X^2$, and $X_3 = X^3$), it is possible to linearize the model and transform it into a multiple-regression problem, which can be solved by the method of least squares. It should be noted that the term linear in the general linear model applies to the dependent variable being a linear function of the unknown parameters. Thus, a general linear model might also include some higher order terms of independent variables, for example, terms such as $X_1^2$, $e^{\beta X}$, $X_1 \cdot X_2$, $1/X$, or $X_2^3$. The basis is, however, to select the proper transformation of input variables or their combinations. Some useful transformations for linearization of the regression model are given in Table 5.3.

The major effort, on the part of a user, in applying multiple-regression techniques lies in identifying the *relevant* independent variables from the initial set and in selecting the regression model using only relevant variables. Two general approaches are common for this task:

1. *Sequential Search Approach.* It is consists primarily of building a regression model with an initial set of variables and then selectively adding or deleting variables until some overall criterion is satisfied or optimized.

2. *Combinatorial Approach.* It is, in essence, a brute-force approach, where the search is performed across all possible combinations of independent variables to determine the best regression model.

Irrespective of whether the sequential or combinatorial approach is used, the maximum benefit to model building occurs from a proper understanding of the application domain.

Additional postprocessing steps may estimate the quality of the linear regression model. Correlation analysis attempts to measure the strength of a relationship between two variables (in our case this relationship is expressed through the linear regression equation). One parameter, which shows this strength of linear association between two variables by means of a single number, is called a *correlation coefficient r*. Its computation requires some intermediate results in a regression analysis.

$$r = \beta\sqrt{(S_{xx} / S_{yy})} = S_{xy} / \sqrt{(S_{xx} \cdot S_{yy})}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \text{mean}_x)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \text{mean}_y)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \text{mean}_x)(y_i - \text{mean}_y)$$

The value of r is between −1 and 1. Negative values for r correspond to regression lines with negative slopes and a positive r shows a positive slope. We must be very careful in interpreting the r value. For example, values of r equal to 0.3 and 0.6 only mean that we have two positive correlations, the second somewhat stronger than the first. It is wrong to conclude that r = 0.6 indicates a linear relationship twice as strong as that indicated by the value r = 0.3.

For our simple example of linear regression given at the beginning of this section, the model obtained was B = 0.8 + 0.92A. We may estimate the quality of the model using the correlation coefficient r as a measure. Based on the available data in Figure 4.3, we obtained intermediate results

$$S_{AA} = 62$$

$$S_{BB} = 60$$

$$S_{AB} = 52$$

and the final correlation coefficient:

$$r = 52 / \sqrt{62 \cdot 60} = 0.85$$

A correlation coefficient r = 0.85 indicates a good linear relationship between two variables. Additional interpretation is possible. Because $r^2 = 0.72$, we can say that approximately 72% of the variations in the values of B is accounted for by a linear relationship with A.

## 5.5 ANOVA

Often the problem of analyzing the quality of the estimated regression line and the influence of the independent variables on the final regression is handled through an ANOVA approach. This is a procedure where the total variation in the dependent variable is subdivided into meaningful components that are then observed and treated in a systematic fashion. ANOVA is a powerful tool that is used in many data-mining applications.

ANOVA is primarily a method of identifying which of the β's in a linear regression model are nonzero. Suppose that the β parameters have already been estimated by the least-square error algorithm. Then the residuals are differences between the observed output values and the fitted values:

$$R_i = y_i - f(x_i)$$

The size of the residuals, for all m samples in a data set, is related to the size of variance $\sigma^2$ and it can be estimated by:

$$S^2 = \left[\sum_{i=1}^{m} (y_i - f(x_i))^2\right] / (m - (n-1))$$

assuming that the model is not over-parametrized. The numerator is called the residual sum while the denominator is called the residual degree of freedom (d.f.).

The key fact about $S^2$ is that it allows us to compare different linear models. If the fitted model is adequate, then $S^2$ is a good estimate of $\sigma^2$. If the fitted model includes redundant terms (some β's are really 0), $S^2$ is still good and close to $\sigma^2$. Only if the fitted model does not include one or more of the inputs that it ought to, will $S^2$ tend to be significantly larger than the true value of $\sigma^2$. These criteria are basic decision steps in the ANOVA algorithm, in which we analyze the influence of input variables on a final model. First, we start with all inputs and compute $S^2$ for this model. Then, we omit inputs from the model one by one. If we omit a useful input the estimate $S^2$ will significantly increase, but if we omit a redundant input the estimate should not change much. Note that omitting one of the inputs from the model is equivalent to forcing the corresponding β to the 0. In principle, in each iteration we compare two $S^2$ values and analyze the differences between them. For this purpose, we introduce an F-ratio or F-statistic test in the form

$$F = S^2_{new} / S^2_{old}$$

If the new model (after removing one or more inputs) is adequate, then F will be close to 1; a value of F significantly larger than one will signal that the model is not

TABLE 5.4. ANOVA for a Data Set with Three Inputs, $x_1$, $x_2$, and $x_3$

| Case | Set of Inputs | $S_i^2$ | F |
|---|---|---|---|
| 1 | $x_1$, $x_2$, $x_3$ | 3.56 | |
| 2 | $x_1$, $x_2$ | 3.98 | $F_{21} = 1.12$ |
| 3 | $x_1$, $x_3$ | 6.22 | $F_{31} = 1.75$ |
| 4 | $x_2$, $x_3$ | 8.34 | $F_{41} = 2.34$ |
| 5 | $x_1$ | 9.02 | $F_{52} = 2.27$ |
| 6 | $x_2$ | 9.89 | $F_{62} = 2.48$ |

adequate. Using this iterative ANOVA approach, we can identify which inputs are related to the output and which are not. The ANOVA procedure is only valid if the models being compared are nested; in other words, one model is a special case of the other.

Suppose that the data set has three input variables, $x_1$, $x_2$, and $x_3$, and one output Y. In preparation for the use of the linear regression method, it is necessary to estimate the simplest model, in terms of the number of required inputs. Suppose that after applying the ANOVA methodology the results given in Table 5.4 are obtained.

The results of ANOVA show that the input attribute $x_3$ does not have an influence on the output estimation because the F-ratio value is close to 1:

$$F_{21} = S_2 / S_1 = 3.98 / 3.56 = 1.12$$

In all other cases, the subsets of inputs increase the F-ratio significantly, and therefore, there is no possibility of reducing the number of input dimensions further without influencing the quality of the model. The final linear regression model for this example will be

$$Y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

*Multivariate* ANOVA (MANOVA) is a generalization of the previously explained ANOVA, and it concerns data-analysis problems in which the output is a vector rather than a single value. One way to analyze this sort of data would be to model each element of the output separately but this ignores the possible relationship between different outputs. In other words, the analysis would be based on the assumption that outputs are not related. MANOVA is a form of analysis that *does* allow correlation between outputs. Given the set of input and output variables, we might be able to analyze the available data set using a multivariate linear model:

$$Y_j = \alpha + \beta_1 \cdot x_{1j} + \beta_2 \cdot x_{2j} + \beta_3 \cdot x_{3j} + \ldots + \beta_n \cdot x_{nj} + \varepsilon_j \quad j = 1, 2, \ldots, m$$

where n is the number of input dimensions, m is the number of samples, $Y_j$ is a vector with dimensions $c \times 1$, and c is the number of outputs. This multivariate model can be fitted in exactly the same way as a linear model using least-square estimation. One way to do this fitting would be to fit a linear model to each of the c dimensions of the output,

one at a time. The corresponding residuals for each dimension will be $(y_j - y'_j)$ where $y_j$ is the exact value for a given dimension and $y'_j$ is the estimated value.

The analog of the residual sum of squares for the univariate linear model is the matrix of the residual sums of squares for the multivariate linear model. This matrix R is defined as

$$R = \sum_{j=1}^{m} (y_j - y_j)(y_j - y_j)^T$$

The matrix R has the residual sum of squares for each of the c dimensions stored on its leading diagonal. The off-diagonal elements are the residual sums of cross-products for pairs of dimensions. If we wish to compare two nested linear models to determine whether certain β's are equal to 0, then we can construct an extra sum of squares matrix and apply a method similar to ANOVA—MANOVA. While we had an F-statistic in the ANOVA methodology, MANOVA is based on matrix R with four commonly used test statistics: Roy's greatest root, the Lawley-Hotteling trace, the Pillai trace, and Wilks' lambda. Computational details of these tests are not explained in the book, but most textbooks on statistics will explain these; also, most standard statistical packages that support MANOVA support all four statistical tests and explain which one to use under what circumstances.

Classical multivariate analysis also includes the method of principal component analysis, where the set of vector samples is transformed into a new set with a reduced number of dimensions. This method has been explained in Chapter 3 when we were talking about data reduction and data transformation as preprocessing phases for data mining.

## 5.6 LOGISTIC REGRESSION

Linear regression is used to model continuous-value functions. Generalized regression models represent the theoretical foundation on that the linear regression approach can be applied to model categorical response variables. A common type of a generalized linear model is *logistic regression*. Logistic regression models the probability of some event occurring as a linear function of a set of predictor variables.

Rather than predicting the value of the dependent variable, the logistic regression method tries to estimate the probability that the dependent variable will have a given value. For example, in place of predicting whether a customer has a good or bad credit rating, the logistic regression approach tries to estimate the probability of a good credit rating. The actual state of the dependent variable is determined by looking at the estimated probability. If the estimated probability is greater than 0.50 then the prediction is closer to *YES* (a good credit rating), otherwise the output is closer to *NO* (a bad credit rating is more probable). Therefore, in logistic regression, the probability p is called the success probability.

We use logistic regression only when the output variable of the model is defined as a categorical binary. On the other hand, there is no special reason why any of the

inputs should not also be quantitative, and, therefore, logistic regression supports a more general input data set. Suppose that output Y has two possible categorical values coded as 0 and 1. Based on the available data we can compute the probabilities for both values for the given input sample: $P(y_j = 0) = 1 - p_j$ and $P(y_j = 1) = p_j$. The model with which we will fit these probabilities is accommodated linear regression:

$$\log(p_j/[1-p_j]) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

This equation is known as the *linear logistic model*. The function $\log(p_j/[1-p_j])$ is often written as *logit(p)*. The main reason for using the logit form of output is to prevent the predicting probabilities from becoming values out of the required range [0, 1]. Suppose that the estimated model, based on a training data set and using the linear regression procedure, is given with a linear equation

$$\text{logit}(p) = 1.5 - 0.6 \cdot x_1 + 0.4 \cdot x_2 - 0.3 \cdot x_3$$

and also suppose that the new sample for classification has input values $\{x_1, x_2, x_3\} = \{1, 0, 1\}$. Using the linear logistic model, it is possible to estimate the probability of the output value 1, (p[Y = 1]) for this sample. First, calculate the corresponding logit(p):

$$\text{logit}(p) = 1.5 - 0.6 \cdot 1 + 0.4 \cdot 0 - 0.3 \cdot 1 = 0.6$$

and then the probability of the output value 1 for the given inputs:

$$\log(p/[1-p]) = 0.6$$

$$p = e^{0.6}/(1+e^{0.6}) = 0.65$$

Based on the final value for probability p, we may conclude that output value Y = 1 is more probable than the other categorical value Y = 0. Even this simple example shows that logistic regression is a very simple yet powerful classification tool in data-mining applications. With one set of data (training set) it is possible to establish the logistic regression model and with other sets of data (testing set) we may analyze the quality of the model in predicting categorical values. The results of logistic regression may be compared with other data-mining methodologies for classification tasks such as decision rules, neural networks, and Bayesian classifier.

## 5.7 LOG-LINEAR MODELS

Log-linear modeling is a way of analyzing the relationship between categorical (or quantitative) variables. The log-linear model approximates discrete, multidimensional probability distributions. It is a type of a generalized linear model where the output Yi is assumed to have a Poisson distribution, with expected value $\mu_j$. The natural logarithm of $\mu_j$ is assumed to be the linear function of inputs

TABLE 5.5. A 2 × 2 Contingency Table for 1100 Samples Surveying Attitudes about Abortion

| | *Support* | | Total |
|---|---|---|---|
| | Yes | No | |
| Sex | | | |
| Female | 309 | 191 | 500 |
| Male | 319 | 281 | 600 |
| Total | 628 | 472 | 1100 |

$$\log(\mu_j) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

Since all the variables of interest are categorical variables, we use a table to represent them, a frequency table that represents the global distribution of data. The aim in log-linear modeling is to identify associations between categorical variables. Association corresponds to the interaction terms in the model, so our problem becomes a problem of finding out which of all β's are 0 in the model. A similar problem can be stated in ANOVA. If there is an interaction between the variables in a log-linear mode, it implies that the variables involved in the interaction are not independent but related, and the corresponding β is not equal to 0. There is no need for one of the categorical variables to be considered as an output in this analysis. If the output is specified, then instead of the log-linear models, we can use logistic regression for analysis. Therefore, we will next explain log-linear analysis when a data set is defined without output variables. All given variables are categorical, and we want to analyze the possible associations between them. That is the task for *correspondence analysis*.

Correspondence analysis represents the set of categorical data for analysis within incidence matrices, also called *contingency tables*. The result of an analysis of the contingency table answers the question: Is there a relationship between analyzed attributes or not? An example of a 2 × 2 contingency table, with cumulative totals, is shown in Table 5.5. The table is a result of a survey to examine the relative attitude of males and females about abortion. The total set of samples is 1100 and every sample consists of two categorical attributes with corresponding values. For the attribute *sex*, the possible values are male and female, and for attribute s*upport* the values are yes and no. Cumulative results for all the samples are represented in four elements of the contingency table.

Are there any differences in the extent of support for abortion between the male and the female populations? This question may be translated to: What is the level of dependency (if any) between the two given attributes: sex and support? If an association exists, then there are significant differences in opinion between the male and the female populations; otherwise both populations have a similar opinion.

Having seen that log-linear modeling is concerned with association of categorical variables, we might attempt to find some quantity (measure) based on this model using data in the contingency table. But we do not do this. Instead, we define the algorithm for feature association based on a comparison of two contingency tables:

1. The first step in the analysis is to transform a given contingency table into a similar table with *expected* values. These expected values are calculated under assumption that the variables are independent.
2. In the second step, we compare these two matrices using the squared distance measure and the chi-square test as criteria of association for two categorical variables.

The computational process for these two steps is very simple for a $2 \times 2$ contingency table. The process is also applicable for increased dimensions of a contingency table (analysis of categorical variables with more than two values, such as $3 \times 4$ or $6 \times 9$).

Let us introduce the notation. Denote the contingency table as $X_{m \times n}$. The row totals for the table are

$$X_{j+} = \sum_{i=1}^{n} X_{ji}$$

and they are valid for every row (j = 1, . . . , m). Similarly, we can define the column totals as

$$X_{+i} = \sum_{j=1}^{m} X_{ji}$$

The grand total is defined as a sum of row totals:

$$X_{++} = \sum_{j=1}^{m} X_{j+}$$

or as a sum of column totals:

$$X_{++} = \sum_{i=1}^{n} X_{+i}$$

Using these totals we can calculate the contingency table of expected values under the assumption that there is no association between the row variable and the column variable. The expected values are

$$E_{ji} = (X_{j+} \cdot X_{+i}) / X_{++} \quad \text{for} \quad j = 1, \ldots, m, \quad i = 1, \ldots, n$$

and they are computed for every position in the contingency table. The final result of this first step will be a totally new table that consists only of expected values, and the two tables will have the same dimensions.

For our example in Table 5.5, all sums (columns, rows, and grand total) are already represented in the contingency table. Based on these values we can construct the

TABLE 5.6. A 2 × 2 Contingency Table of Expected Values for the Data Given in Table 5.5

| | *Support* | | Total |
|---|---|---|---|
| | Yes | No | |
| Sex | | | |
| Female | 285.5 | 214.5 | 500 |
| Male | 342.5 | 257.5 | 600 |
| Total | 628 | 472 | 1100 |

contingency table of expected values. The expected value on the intersection of the first row and the first column will be

$$E_{11} = (X_{1+} \cdot X_{+1}) / X_{++} = 500 \cdot 628 / 1100 = 285.5$$

Similarly, we can compute the other expected values and the final contingency table with expected values will be as given in Table 5.6.

The next step in the analysis of categorical-attributes dependency is the application of the chi-squared test of association. The initial hypothesis $H_0$ is the assumption that the two attributes are unrelated, and it is tested by Pearson's chi-squared formula:

$$\chi^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} ((X_{ji} - E_{ji})^2 / E_{ji})$$

The greater the value of $\chi^2$, the greater the evidence against the hypothesis $H_0$ is. For our example, comparing Tables 5.5 and 5.6, the test gives the following result:

$$\chi^2 = 8.2816$$

with the d.f. for an m × n dimensional table computed as

$$\text{d.f.} = (m-1) \cdot (n-1) = (2-1)(2-1) = 1$$

In general, the hypothesis $H_0$ is rejected at the level of significance $\alpha$ if

$$\chi^2 \geq T(\alpha)$$

where $T(\alpha)$ is the threshold value from the $\chi^2$ distribution table usually given in textbooks on statistics. For our example, selecting $\alpha = 0.05$ we obtain the threshold

$$T(0.05) = \chi^2(1-\alpha, \text{d.f.}) = \chi^2(0.95, 1) = 3.84.$$

A simple comparison shows that

$$\chi^2 = 8.2816 \geq T(0.05) = 3.84$$

TABLE 5.7. Contingency Tables for Categorical Attributes with Three Values

*(a) A 3 × 3 contingency table of observed values*

| Attribute1 | | Low | Medium | High | Total |
|---|---|---|---|---|---|
| Attribute2 | Excellent | 21 | 11 | 4 | 36 |
| | Good | 3 | 2 | 2 | 7 |
| | Poor | 7 | 1 | 1 | 9 |
| Total | | 31 | 14 | 7 | 52 |

*(b) A 3 × 3 contingency table of expected values under $H_0$*

| Attribute1 | | Low | Medium | High | Total |
|---|---|---|---|---|---|
| Attribute2 | Excellent | 21.5 | 9.7 | 4.8 | 36 |
| | Good | 4.2 | 1.9 | 0.9 | 7 |
| | Poor | 5.4 | 2.4 | 1.2 | 9 |
| Total | | 31 | 14 | 7 | 52 |

and therefore, we can conclude that hypothesis $H_0$ is rejected; the attributes analyzed in the survey have a high level of dependency. In other words, the attitude about abortion shows differences between the male and the female populations.

The same procedure may be generalized and applied to contingency tables where the categorical attributes have more than two values. The next example shows how the previously explained procedure can be applied without modifications to the contingency table 3 × 3. The values given in Table 5.7a are compared with the estimated values given in Table 5.7b, and the corresponding test is calculated as $\chi^2 = 3.229$. Note that in this case parameter

$$\text{d.f.} = (n-1)(m-1) = (3-1)\cdot(3-1) = 4.$$

We have to be very careful about drawing additional conclusions and further analyzing the given data set. It is quite obvious that the sample size is not large. The number of observations in many cells of the table is small. This is a serious problem and additional statistical analysis is necessary to check if the sample is a good representation of the total population or not. We do not cover this analysis here because in most real-world data-mining problems the data set is enough large to eliminate the possibility of occurrence of these deficiencies.

That was one level of generalization for an analysis of contingency tables with categorical data. The other direction of generalization is inclusion into analysis of more than two categorical attributes. The methods for three- and high-dimensional contingency table analysis are described in many books on advanced statistics; they explain the procedure of discovered dependencies between several attributes that are analyzed simultaneously.

## 5.8 LDA

LDA is concerned with classification problems where the dependent variable is categorical (nominal or ordinal) and the independent variables are metric. The objective of

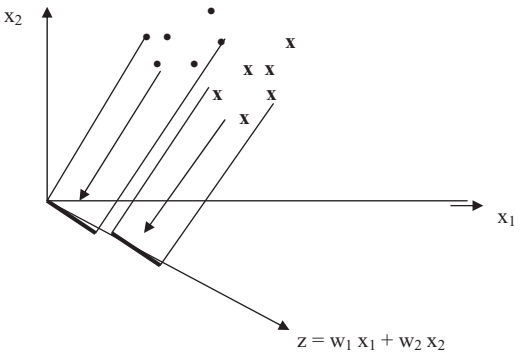


Figure 5.5. Geometric interpretation of the discriminant score.

LDA is to construct a discriminant function that yields different scores when computed with data from different output classes. A linear discriminant function has the following form:

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_k x_k$$

where $x_1, x_2, \ldots, x_k$ are independent variables. The quantity z is called the discriminant score, and $w_1, w_2, \ldots, w_k$ are called weights. A geometric interpretation of the discriminant score is shown in Figure 5.5. As the figure shows, the discriminant score for a data sample represents its projection onto a line defined by the set of weight parameters.

The construction of a discriminant function z involves finding a set of weight values $w_i$ that maximizes the ratio of the *between-class* to the *within-class* variance of the discriminant score for a preclassified set of samples. Once constructed, the discriminant function z is used to predict the class of a new nonclassified sample. Cutting scores serve as the criteria against which each individual discriminant score is judged. The choice of cutting scores depends upon a distribution of samples in classes. Letting $z_a$ and $z_b$ be the mean discriminant scores of preclassified samples from class A and B, respectively, the optimal choice for the cutting score $z_{cut\text{-}ab}$ is given as

$$z_{cut\text{-}ab} = (z_a + z_b)/2$$

when the two classes of samples are of equal size and are distributed with uniform variance. A new sample will be classified to one or another class depending on its score $z > z_{cut\text{-}ab}$ or $z < z_{cut\text{-}ab}$. A weighted average of mean discriminant scores is used as an optimal cutting score when the set of samples for each of the classes are not of equal size:

$$z_{cut\text{-}ab} = (n_a \cdot z_a + n_b \cdot z_b)/(n_a + n_b)$$

The quantities $n_a$ and $n_b$ represent the number of samples in each class. Although a single discriminant function z with several discriminant cuts can separate samples


Figure 5.6. Classification process in multiple-discriminant analysis.

into several classes, *multiple discriminant analysis* is used for more complex problems. The term multiple discriminant analysis is used in situations when separate discriminant functions are constructed for each class. The classification rule in such situations takes the following form: Decide in favor of the class whose discriminant score is the highest. This is illustrated in Figure 5.6.

## 5.9 REVIEW QUESTIONS AND PROBLEMS

**1.** What are the differences between statistical testing and estimation as basic areas in statistical inference theory?

**2.** A data set for analysis includes only one attribute X:

X = {7, 12, 5, 18, 5, 9, 13, 12, 19, 7, 12, 12, 13, 3, 4, 5, 13, 8, 7, 6}.

(a) What is the mean of the data set X?
(b) What is the median?
(c) What is the mode, and what is the modality of the data set X?
(d) Find the standard deviation for X.
(e) Give a graphical summarization of the data set X using a boxplot representation.
(f) Find outliers in the data set X. Discuss the results.

**3.** For the training set given in Table 5.1, predict the classification of the following samples using simple Bayesian classifier.

(a) {2, 1, 1}
(b) {0, 1, 1}

**4.** Given a data set with two dimensions X and Y:

| X | Y |
|---|---|
| 1 | 5 |
| 4 | 2.75 |
| 3 | 3 |
| 5 | 2.5 |

(a) Use a linear regression method to calculate the parameters $\alpha$ and $\beta$ where $y = \alpha + \beta x$.
(b) Estimate the quality of the model obtained in (a) using the correlation coefficient r.
(c) Use an appropriate nonlinear transformation (one of those represented in Table 5.3) to improve regression results. What is the equation for a new, improved, and nonlinear model? Discuss a reduction of the correlation coefficient value.

**5.** A logit function, obtained through logistic regression, has the form:

$$\text{Logit}(p) = 1.2 - 1.3\,x1 + 0.6\,x2 + 0.4\,x3$$

Find the probability of output values 0 and 1 for the following samples:

(a) { 1, −1, −1 }
(b) { −1, 1, 0 }
(c) { 0, 0, 0 }

**6.** Analyze the dependency between categorical attributes X and Y if the data set is summarized in a $2 \times 3$ contingency table:

| | | Y | |
|---|---|---|---|
| | | T | F |
| X | A | 128 | 7 |
| | B | 66 | 30 |
| | C | 42 | 55 |

**7.** Implement the algorithm for a boxplot representation of each numeric attribute in an input flat file.

**8.** What are the basic principles in the construction of a discriminant function applied in an LDA?

**9.** Implement the algorithm to analyze a dependency between categorical variables using two-dimensional contingency tables.

**10.** Find *EMA*(4, 4) for the data set {27, 27, 18, 9} if: (a) p = 1/3, and (b) p = 3/4. Discuss the solutions.

**11.** With Bayes classifier, missing data items are

(a) treated as equal compares
(b) treated as unequal compares
(c) replaced with a default value
(d) ignored

Determine what is the true statement.

**12.** The table below contains counts and ratios for a set of data instances to be used for supervised Bayesian learning. The output attribute is sex with possible values *male* and *female.* Consider an individual who has said *no* to the life insurance promotion, *yes* to the magazine promotion, *yes* to the watch promotion, and has credit card insurance. Use the values in the table together with Bayes classifier to determine the probability that this individual is *male.*

| | *Magazine Promotion* | | *Watch Promotion* | | *Life Insurance Promotion* | | *Credit Card Insurance* | |
|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female |
| Yes | 4 | 3 | 2 | 2 | 2 | 3 | 2 | 1 |
| No | 2 | 1 | 4 | 2 | 4 | 1 | 4 | 3 |

**13.** The probability that a person owns a sports car given that they subscribe to at least one automotive magazine is 40%. We also know that 3% of the adult population subscribes to at least one automotive magazine. Finally, the probability of a person owning a sports car given that they do not subscribe to at least one automotive magazine is 30%. Use this information together with Bayes theorem to compute the probability that a person subscribes to at least one automotive magazine given that they own a sports car.

**14.** Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

**15.** Given a $2 \times 2$ contingency table for X and Y attributes:

| | | X | |
|---|---|---|---|
| | | x1 | x2 |
| Y | y1 | 7 | 4 |
| | y2 | 2 | 8 |

(a) Find a contingency table with *expected values.*
(b) If the threshold value for $\chi^2$ test is 8.28, determine if attributes X and Y are dependent or not.

**16.** The logit function, obtained through logistic regression, has a form:

$$\text{Logit}(p) = 1.2 - 1.3 \times 1 + 0.6 \times 2 + 0.4 \times 3$$

Find the probability of output values 0 and 1 for the sample $\{1, -1, -1\}$.

**17.** Given:

- P(Good Movie | Includes Tom Cruise) = 0.01
- P(Good Movie | Tom Cruise absent) = 0.1
- P(Tom Cruise in a randomly chosen movie) = 0.01

What is P(Tom Cruise is in the movie | Not a Good Movie)?

**18.** You have the following training set with three Boolean input x, y, and z, and a Boolean output U. Suppose you have to predict U using a naive Bayesian classifier.

| x | y | z | U |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 |

(a) After learning is complete what would be the predicted probability P(U = 0|x = 0, y = 1, z = 0)?

(b) Using the probabilities obtained during the Bayesian classifier training, what would be the predicted probability P(U = 0|x = 0)?

## 5.10 REFERENCES FOR FURTHER STUDY

Berthold, M., D. J. Hand, eds., *Intelligent Data Analysis—An Introduction*, Springer, Berlin, 1999.

The book is a detailed introductory presentation of the key classes of intelligent data-analysis methods including all common data-mining techniques. The first half of the book is devoted to the discussion of classical statistical issues, ranging from basic concepts of probability and inference to advanced multivariate analyses and Bayesian methods. The second part of the book covers theoretical explanations of data-mining techniques with their roots in disciplines other than statistics. Numerous illustrations and examples will enhance a reader's knowledge about the theory and practical evaluations of data-mining techniques.

Brandt, S., *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, 3rd edition, Springer, New York, 1999.

This text bridges the gap between statistical theory and practice. It emphasizes concise but rigorous mathematics while retaining the focus on applications. After introducing probability and random variables, the book turns to the generation of random numbers and important distributions. Subsequent chapters discuss statistical samples, the maximum-likelihood method, and the testing of statistical hypotheses. The text concludes with a detailed discussion of several important statistical methods such as least-square minimization, ANOVA, regressions, and analysis of time series.

Cherkassky, V., F. Mulier, *Learning from Data: Concepts, Theory and Methods*, John Wiley, New York, 1998.

The book provides a unified treatment of the principles and methods for learning dependencies from data. It establishes a general conceptual framework in which various learning methods from statistics, machine learning, and other disciplines can be applied—showing that a few fundamental principles underlie most new methods being proposed today. An additional

strength of this primary theoretical book is a large number of case studies and examples that simplify and make understandable statistical learning theory concepts.

Hand, D., Mannila H., Smith P., *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.

The book consists of three sections. The first, foundations, provides a tutorial overview of the principles underlying data-mining algorithms and their applications. The second section, data-mining algorithms, shows how algorithms are constructed to solve specific problems in a principled manner. The third section shows how all of the preceding analyses fit together when applied to real-world data-mining problems.

Nisbet, R., J. Elder, G. Miner, *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier Inc., Amsterdam, 2009.

The book is a comprehensive professional reference book that guides business analysts, scientists, engineers, and researchers (both academic and industrial) through all stages of data analysis, model building, and implementation. The handbook helps one discern technical and business problems, understand the strengths and weaknesses of modern data-mining algorithms, and employ the right statistical methods for practical application. Use this book to address massive and complex data sets with novel statistical approaches and be able to objectively evaluate analyses and solutions. It has clear, intuitive explanations of the principles and tools for solving problems using modern analytic techniques, and discusses their application to real problems, in ways accessible and beneficial to practitioners across industries—from science and engineering, to medicine, academia, and commerce. This handbook brings together, in a single resource, all the information a beginner will need to understand the tools and issues in data mining to build successful data-mining solutions.