

## Data Mining: A Closer Look

### Chapter Objectives

- ▶ Determine an appropriate data mining strategy for a specific problem.
- ▶ Know about several data mining techniques and how each technique builds a generalized model to represent data.
- ▶ Understand how a confusion matrix is used to help evaluate supervised learner models.
- ▶ Understand basic techniques for evaluating supervised learner models with numeric output.
- ▶ Know how measuring lift can be used to compare the performance of several competing supervised learner models.
- ▶ Understand basic techniques for evaluating unsupervised learner models.

Although the field of data mining is in a continual state of change, a few basic strategies have remained constant. In Section 2.1 we define five fundamental data mining strategies and give examples of problems appropriate for each strategy. Whereas a data mining strategy outlines an approach for problem solution, a data mining technique applies a strategy. In Sections 2.2 through 2.4 we introduce several data mining techniques with the help of a hypothetical database containing customer information about credit card promotions. Section 2.2 is dedicated to supervised learning techniques. In Section 2.3 we present an overview of association rules, leaving a more detailed discussion for Chapter 3. In Section 2.4 we discuss unsupervised clustering. As you saw in Chapter 1, evaluation is a fundamental step in the data mining process. Section 2.5 provides a few basic tools to help you better understand the evaluation process.

## 2.1 Data Mining Strategies

---

As you learned in Chapter 1, **data mining strategies** can be broadly classified as either supervised or unsupervised. Supervised learning builds models by using input attributes to predict output attribute values. Many supervised data mining algorithms only permit a single output attribute. Other supervised learning tools allow us to specify one or several output attributes. Output attributes are also known as **dependent variables** as their outcome depends on the values of one or more input attributes. Input attributes are referred to as **independent variables**. When learning is unsupervised, an output attribute does not exist. Therefore all attributes used for model building are independent variables.

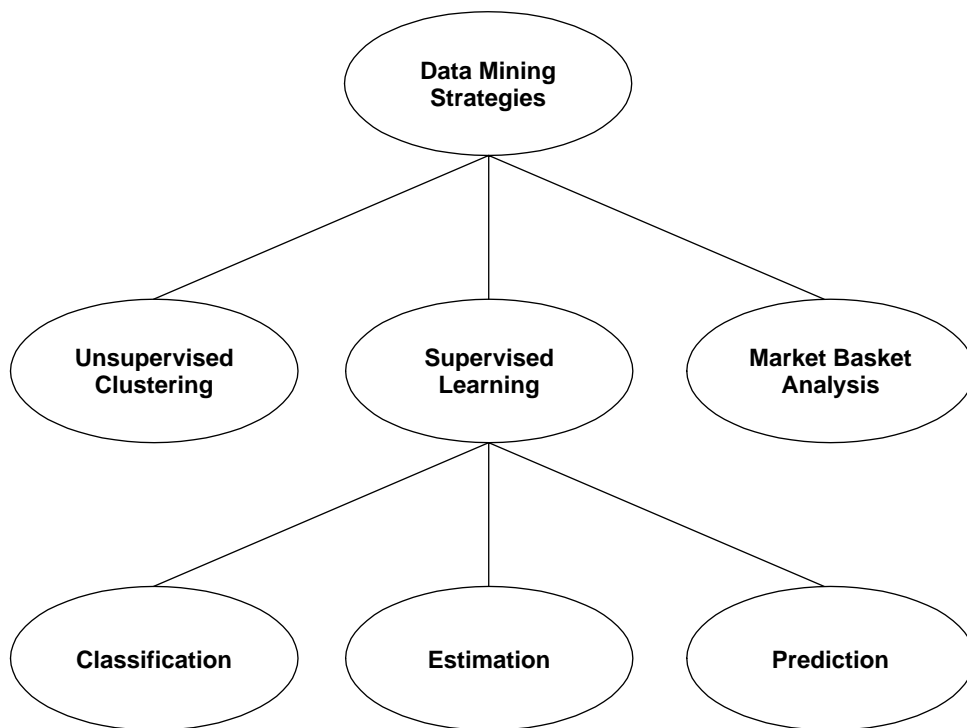
Supervised learning strategies can be further labeled according to whether output attributes are discrete or categorical, as well as by whether models are designed to determine a current condition or predict future outcome. In this section we examine three supervised learning strategies, take a closer look at unsupervised clustering, and introduce a strategy for discovering associations among retail items sold in catalogs and stores. Figure 2.1 shows the five data mining strategies we will discuss.

### Classification

**Classification** is probably the best understood of all data mining strategies. Classification tasks have three common characteristics:

- Learning is supervised.
- The dependent variable is categorical.

Figure 2.1 • **A hierarchy of data mining strategies**



- The emphasis is on building models able to assign new instances to one of a set of well-defined classes.

Some example classification tasks include the following:

- Determine those characteristics that differentiate individuals who have suffered a heart attack from those who have not.
- Develop a profile of a “successful” person.
- Determine if a credit card purchase is fraudulent.
- Classify a car loan applicant as a good or a poor credit risk.
- Develop a profile to differentiate female and male stroke victims.

Notice that each example deals with current rather than future behavior. For example, we want the car loan application model to determine whether an applicant is a good credit risk at this time rather than in some future time period. Prediction models are designed to answer questions about future behavior. Prediction models are discussed in Section 2.1.3.

## Estimation

Like classification, the purpose of an **estimation** model is to determine a value for an unknown output attribute. However, unlike classification, the output attribute for an estimation problem is numeric rather than categorical. Here are four examples of estimation tasks:

- Estimate the number of minutes before a thunderstorm will reach a given location.
- Estimate the salary of an individual who owns a sports car.
- Estimate the likelihood that a credit card has been stolen.
- Estimate the length of a gamma ray burst.

Most supervised data mining techniques are able to solve classification or estimation problems, but not both. If our data mining tool supports one strategy but not the other, we can usually adapt a problem for solution by either strategy. To illustrate, suppose the output attribute for the original training data in the stolen credit card example above is numeric. Let's also assume the output values range between 0 and 1, with 1 being a most likely case for a stolen card. We can make discrete categories for the output attribute values by replacing scores ranging between 0.0 and 0.3 with the value *unlikely*, scores between 0.3 and 0.7 with *likely*, and scores greater than 0.7 with *highly likely*. In this case the transformation between numeric values and discrete categories is straightforward. Cases such as attempting to make monetary amounts discrete present more of a challenge.

## Prediction

It is not easy to differentiate prediction from classification or estimation. However, unlike a classification or estimation model, the purpose of a predictive model is to determine future outcome rather than current behavior. The output attribute(s) of a predictive model can be categorical or numeric. Here are several examples of tasks appropriate for predictive data mining:

- Predict the total number of touchdowns an NFL running back will score during the 2002 NFL season.

- Determine whether a credit card customer is likely to take advantage of a special offer made available with their credit card billing.
- Predict next week's closing price for the Dow Jones Industrial Average.
- Forecast which telephone subscribers are likely to change providers during the next three months.

Most supervised data mining techniques appropriate for classification or estimation problems can also build predictive models. Actually, it is the nature of the data that determines whether a model is suitable for classification, estimation, or prediction. To show this, let's consider a real medical dataset with 303 instances. One hundred sixty-five instances hold information about patients who are free of heart disease. The remaining 138 instances contain data about patients who have a known heart condition.

The attributes and possible attribute values associated with this dataset are shown in Table 2.1. Two forms of the dataset exist. One dataset consists of all numeric attributes. The second dataset has categorical conversions for seven of the original numeric attributes. The table column labeled *Mixed Values* shows the value *Numeric* for attributes that were not converted to a categorical equivalent. For example, the values for attribute *Age* are identical for both datasets. However, the attribute *Fasting Blood Sugar*

### The Cardiology Patient Data Set

The cardiology patient dataset is part of the dataset package that comes with your iDA software. The original data was gathered by Dr. Robert Detrano at the VA Medical Center in Long Beach, California. The dataset consists of 303 instances. One hundred thirty-eight of the instances hold information about patients with heart disease. The original dataset contains 13 numeric attributes and a fourteenth attribute indicating whether the patient has a heart condition. The dataset was later modified by Dr. John Gennari. He changed seven of the numerical attributes to categorical equivalents for the pur-

pose of testing data mining tools able to classify datasets with mixed data types. The Microsoft Excel file names for the datasets are *CardiologyNumerical.xls* and *CardiologyCategorical.xls*, respectively. This dataset is interesting because it represents real patient data and has been used extensively for testing various data mining techniques. We can use this data together with one or more data mining techniques to help us develop profiles for differentiating individuals with heart disease from those without known heart conditions.



Table 2.1 • **Cardiology Patient Data**

Attribute Name	Mixed Values	Numeric Values	Comments
Age	Numeric	*	Age in years
Sex	Male, Female	1, 0	Patient gender (1, 0)
Chest Pain Type	Angina, Abnormal Angina, NoTang, Asymptomatic	1–4	NoTang = Nonanginal pain
Blood Pressure	Numeric	*	Resting blood pressure upon hospital admission
Cholesterol	Numeric	*	Serum cholesterol
Fasting Blood Sugar < 120	True, False	1, 0	Is fasting blood sugar less than 120?
Resting ECG	Normal, Abnormal, Hyp	0, 1, 2	Hyp = Left ventricular hypertrophy
Maximum Heart Rate	Numeric	*	Maximum heart rate achieved
Induced Angina?	True, False	1, 0	Does the patient experience angina as a result of exercise?
Old Peak	Numeric	*	ST depression induced by exercise relative to rest
Slope	Up, flat, down	1–3	Slope of the peak exercise ST segment
Number Colored Vessels	0, 1, 2, 3	0, 1, 2, 3	Number of major vessels colored by fluoroscopy
Thal	Normal fix, rev	3, 6, 7	Normal, fixed defect, reversible defect
Concept Class	Healthy, Sick	1, 0	Angiographic disease status

Table 2.2 • Most and Least Typical Instances from the Cardiology Domain

Attribute Name	Most Typical Healthy Class	Least Typical Healthy Class	Most Typical Sick Class	Least Typical Sick Class
Age	52	63	60	62
Sex	Male	Male	Male	Female
Chest Pain Type	NoTang	Angina	Asymptomatic	Asymptomatic
Blood Pressure	138	145	125	160
Cholesterol	223	233	258	164
Fasting Blood Sugar < 120	False	True	False	False
Resting ECG	Normal	Hyp	Hyp	Hyp
Maximum Heart Rate	169	150	141	145
Induced Angina?	False	False	True	False
Old Peak	0	2.3	2.8	6.2
Slope	Up	Down	Flat	Down
Number of Colored Vessels	0	0	1	3
Thal	Normal	Fix	Rev	Rev

<120 has values *True* or *False* in the converted dataset and values 0 and 1 in the original data.

Table 2.2 lists four instances from the mixed form of the dataset. Two of the instances represent the most typical exemplars from each respective class. The remaining two instances are atypical class members. Some differences between the most typical healthy and the most typical sick patient are easily anticipated. This is the case with typical healthy and sick class values for *Resting ECG* and *Induced Angina*. Surprisingly, we do not see expected differences in cholesterol and blood pressure readings between healthy and sick individuals.

Here are two rules generated for this data by a production rule generator. *Concept class* is specified as the output attribute:

```

IF 169 <= Maximum Heart Rate <= 202
THEN Concept Class = Healthy
      Rule accuracy: 85.07%
      Rule coverage: 34.55%

```

**IF** *Thal = Rev and Chest Pain Type = Asymptomatic*

**THEN** *Concept Class = Sick*

Rule accuracy: 91.14%

Rule coverage: 52.17%

For the first rule the rule accuracy tells us that if a patient has a maximum heart rate between 169 and 202, we will be correct more than 85 times out of 100 in identifying the patient as healthy. Rule coverage reveals that over 34 percent of all healthy patients have a maximum heart rate in the specified range. When we combine this knowledge with the maximum heart rate values shown in Table 2.2, we are able to conclude that healthy patients are likely to have higher maximum heart rate values.

Is this first rule appropriate for classification or prediction? If the rule is predictive, we can use the rule to warn healthy folks with the statement:

**WARNING 1:** Have your maximum heart rate checked on a regular basis. If your maximum heart rate is low, you may be at risk of having a heart attack!

If the rule is appropriate for classification but not prediction, the scenario reads:

**WARNING 2:** If you have a heart attack, expect your maximum heart rate to decrease.

In any case, we cannot imply the stronger statement:

**WARNING 3:** A low maximum heart rate will cause you to have a heart attack!

That is, with data mining we can state relationships between attributes but we cannot say whether the relationships imply causality. Therefore entertaining an exercise program to increase maximum heart rate may or may not be a good idea.

The question still remains as to whether either of the first two warnings are correct. This question is not easily answered. A data mining specialist can develop models to generate rules such as those just given. Beyond this, the specialist must have access to additional information—in this case a medical expert—before determining how to use discovered knowledge.

## Unsupervised Clustering

With unsupervised clustering we are without a dependent variable to guide the learning process. Rather, the learning program builds a knowledge structure by using some measure of cluster quality to group instances into two or more classes. A primary goal of an unsupervised clustering strategy is to discover concept structures in data. Common uses of unsupervised clustering include:

- Determine if meaningful relationships in the form of concepts can be found in the data
- Evaluate the likely performance of a supervised learner model
- Determine a best set of input attributes for supervised learning
- Detect outliers

You saw an obvious use of unsupervised clustering in Chapter 1 when we showed how clustering was applied to the Acme Investors database to find interesting relationships in the form of concept classes in the data. However, it is not unusual to use unsupervised clustering as an evaluation tool for supervised learning.

To illustrate this idea, let's suppose we have built a supervised learner model using the heart patient data with output attribute *Concept Class*. To evaluate the supervised model, we present the training instances to an unsupervised clustering system. The attribute *Concept Class* is flagged as unused. Next, we examine the output of the unsupervised model to determine if the instances from each concept class (*Healthy* and *Sick*) naturally cluster together. If the instances from the individual classes do not cluster together, we may conclude that the attributes are unable to distinguish healthy patients from those with a heart condition. This being the case, the supervised model is likely to perform poorly. One solution is to revisit the attribute and instance choices used to create the supervised model. In fact, choosing a best set of attributes for a supervised learner model can be implemented by repeatedly applying unsupervised clustering with alternative attribute choices. In this way, those attributes best able to differentiate the classes known to be present in the data can be determined. Unfortunately, even with a small number of attribute choices, the application of this technique can be computationally unmanageable.

Unsupervised clustering can also help detect any atypical instances present in the data. Atypical instances are referred to as **outliers**. Outliers can be of great importance and should be identified whenever possible. Statistical mining applications frequently remove outliers. With data mining, the outliers might be just those instances we are trying to identify. For example, an application that checks credit card purchases would likely identify an outlier as a positive instance of credit card fraud. One way to find outliers is to perform an unsupervised clustering and examine those instances that do not group naturally with the other instances.

## Market Basket Analysis

The purpose of **market basket analysis** is to find interesting relationships among retail products. The results of a market basket analysis help retailers design promotions, arrange shelf or catalog items, and develop cross-marketing strategies. Association rule algorithms are often used to apply a market basket analysis to a set of data. Association rules are briefly described later in this chapter and are presented in detail in Chapter 3.

## 2.2 Supervised Data Mining Techniques

---

A **data mining technique** is used to apply a data mining strategy to a set of data. A specific data mining technique is defined by an algorithm and an associated knowledge structure such as a tree or a set of rules. In Chapter 1 we introduced decision trees as the most studied of all supervised data mining techniques. Here we present several additional supervised data mining methods. Our goal is to help you develop a basic understanding of the similarities and differences between the various data mining techniques.

### The Credit Card Promotion Database

We will use the fictitious data displayed in Table 2.3 to help explain the data mining methods presented here. The table shows data extracted from a database containing in-

#### The Credit Card Promotion Database

Credit card companies often include promotional offerings with their monthly credit card billings. The offers provide the credit card customer with an opportunity to purchase items such as luggage, magazines, or jewelry. Credit card companies sponsoring new promotions frequently send bills to individuals without a current card balance hoping that some of these individuals will take advantage of one or more of the promotional offerings. From the perspective of predictive data mining, given the right data, we may be able to find relationships that provide insight about the characteristics of individuals likely to take advantage of future promotions. In doing so, we can divide the pool of zero-balance card holders into two classes. One class will be those persons likely to take advantage of a new credit card promotion. These individuals should be sent a zero-balance billing

containing the promotional information. The second class will consist of persons not likely to make a promotional purchase. These individuals should not be sent a zero-balance monthly statement. The end result is a savings in the form of decreased postage, paper, and processing costs for the credit card company.

The credit card promotion database shown in Table 2.3 has fictitious data about 15 individuals holding credit cards with the Acme Credit Card Company. The data contains information obtained about customers through their initial credit card application as well as data about whether these individuals have accepted various promotional offerings sponsored by the credit card company. Although the dataset is small, it serves well for purposes of illustration. We employ this dataset for descriptive purposes throughout the text. ■

formation collected on individuals who hold credit cards issued by the Acme Credit Card Company. The first row of Table 2.3 contains the attribute names for each column of data. The first column gives the salary range for an individual credit card holder. Values in columns two through four tell us which card holders have taken advantage of specified promotions sent with their monthly credit card bill. Column five tells us whether an individual has credit card insurance. Column six gives the gender of the card holder, and column seven offers the card holder's age. The first card holder shown in the table has a yearly salary between \$40,000 and \$50,000, is a 45-year-old male, has purchased one or several magazines advertised with one of his credit card bills, did not take advantage of any other credit card promotions, and does not have credit card insurance. Several attributes likely to be relevant for data mining purposes are not included in the table. Some of these attributes are promotion dates, dollar amounts for purchases, average monthly credit card balance, and marital status. Let's turn our attention to the data mining techniques to see what they can find in the credit card promotion database.

## Production Rules

In Chapter 1 you saw that any decision tree can be translated into a set of production rules. However, we do not need an initial tree structure to generate production rules. **RuleMaker**, the production rule generator that comes with your iDA software, uses ratios together with mathematical set theory operations to create rules from spreadsheet data. Earlier in this chapter you saw two rules generated by RuleMaker for the heart patient dataset. Let's apply RuleMaker to the credit card promotion data.

For our experiment we will assume the Acme Credit Card Company has authorized a new life insurance promotion similar to the previous promotion specified in Table 2.3. The promotion material will be sent as part of the credit card billing for all card holders with a non-zero balance. We will use data mining to help us send billings to a select group of individuals who do not have a current credit card balance but are likely to take advantage of the promotion.

Our problem calls for supervised data mining using *life insurance promotion* as the output attribute. Our goal is to develop a profile for individuals likely to take advantage of a life insurance promotion advertised along with their next credit card statement. Here is a possible hypothesis:

*A combination of one or more of the dataset attributes differentiate between Acme Credit Card Company card holders who have taken advantage of a life insurance promotion and those card holders who have chosen not to participate in the promotional offer.*

Table 2.3 • The Credit Card Promotion Database

Income Range (\$)	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	Yes	No	No	No	Male	45
30–40K	Yes	Yes	Yes	No	Female	40
40–50K	No	No	No	No	Male	42
30–40K	Yes	Yes	Yes	Yes	Male	43
50–60K	Yes	No	Yes	No	Female	38
20–30K	No	No	No	No	Female	55
30–40K	Yes	No	Yes	Yes	Male	35
20–30K	No	Yes	No	No	Male	27
30–40K	Yes	No	No	No	Male	43
30–40K	Yes	Yes	Yes	No	Female	41
40–50K	No	Yes	Yes	No	Female	43
20–30K	No	Yes	Yes	No	Male	29
50–60K	Yes	Yes	Yes	No	Female	39
40–50K	No	Yes	No	No	Male	55
20–30K	No	No	Yes	Yes	Female	19

The hypothesis is stated in terms of current rather than predicted behavior. However, the nature of the created rules will tell us whether we can use the rules for classification or prediction.

When presented with this data, the iDA rule generator offered several rules of interest. Here are four such rules:

- IF** *Sex = Female & 19 ≤ Age ≤ 43*  
**THEN** *Life Insurance Promotion = Yes*  
 Rule Accuracy: 100.00%  
 Rule Coverage: 66.67%
- IF** *Sex = Male & Income Range = 40–50K*  
**THEN** *Life Insurance Promotion = No*  
 Rule Accuracy: 100.00%  
 Rule Coverage: 50.00%

3. **IF** *Credit Card Insurance = Yes*  
**THEN** *Life Insurance Promotion = Yes*  
Rule Accuracy: 100.00%  
Rule Coverage: 33.33%
  
4. **IF** *Income Range = 30–40K & Watch Promotion = Yes*  
**THEN** *Life Insurance Promotion = Yes*  
Rule Accuracy: 100.00%  
Rule Coverage: 33.33%

The first rule tells us that we should send a credit card bill containing the promotion to all females between the ages of 19 and 43. Although the coverage for this rule is 66.67%, it would be too optimistic to assume that two-thirds of all females in the specified age range will take advantage of the promotion. The second rule indicates that males who make between \$40,000 and \$50,000 a year are not good candidates for the insurance promotion. The 100.00% accuracy tells us that our sample does not contain a single male within the \$40,000 to \$50,000 income range who took advantage of the previous life insurance promotion.

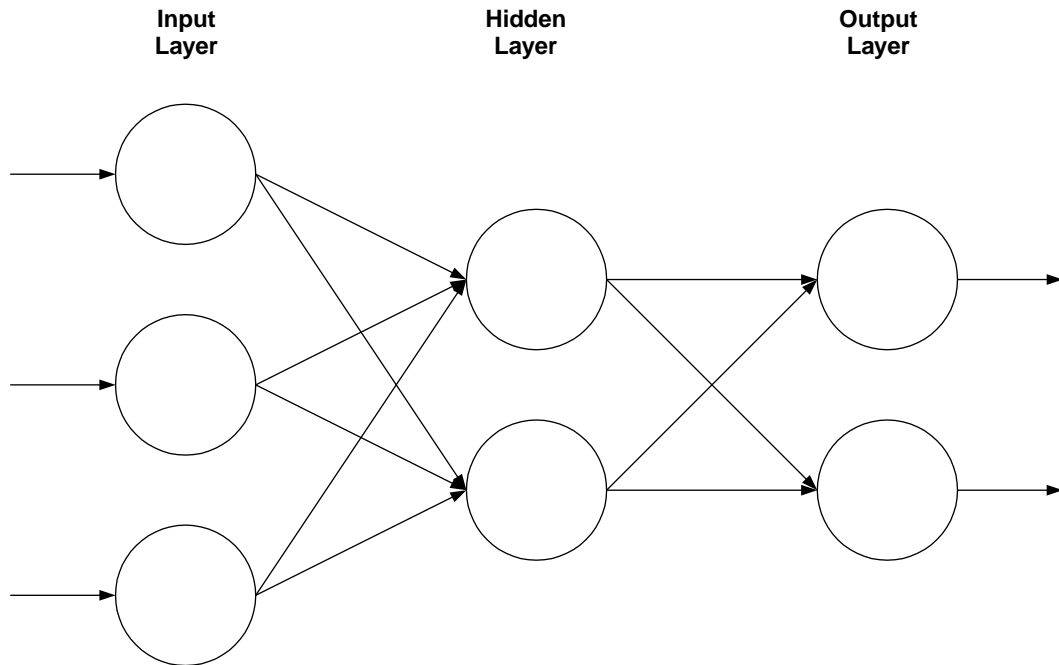
The first and second rules are particularly helpful as neither rule contains an antecedent condition involving a previous promotion. The rule preconditions are based purely on information obtained at the time of initial application. As credit card insurance is always initially offered upon a card approval, the third rule is also useful. However, the fourth rule will not be applicable to new card holders who have not had a chance to take advantage of a previous promotion. For new card holders we should consider the first three rules as predictive and the fourth rule as effective for classification but not predictive purposes.

## Neural Networks

A **neural network** is a set of interconnected nodes designed to imitate the functioning of the human brain. As the human brain contains billions of neurons and a typical neural network has fewer than one hundred nodes, the comparison is somewhat superficial. However, neural networks have been successfully applied to problems across several disciplines and for this reason are quite popular in the data mining community.

Neural networks come in many shapes and forms and can be constructed for supervised learning as well as unsupervised clustering. In all cases the values input into a neural network must be numeric. The feed-forward network is a popular supervised learner model. Figure 2.2 shows a fully connected feed-forward neural network consisting of three layers. With a feed-forward network the input attribute values for an individual instance enter at the input layer and pass directly through the output layer

Figure 2.2 • A multilayer fully connected neural network



of the network structure. The output layer may contain one or several nodes. The output layer of the network shown in Fig. 2.2 contains two nodes. Therefore the output of the neural network will be an ordered pair of values.

The network displayed in Fig. 2.2 is fully connected, as the nodes at one layer are connected to all nodes at the next layer. In addition, each network node connection has an associated weight (not shown in the diagram). Notice that nodes within the same layer of the network architecture are not connected to one another.

Neural networks operate in two phases. The first phase is called the learning phase. During network learning, the input values associated with each instance enter the network at the input layer. One input layer node exists for each input attribute contained in the data. The actual output value for each instance is computed and compared with the desired network output. Any error between the desired and computed output is propagated back through the network by changing connection-weight values. Training terminates after a certain number of iterations or when the network converges to a predetermined minimum error rate. During the second phase of operation, the network weights are fixed and the network is used to classify new instances.

Your iDA software suite of tools contains a feed-forward neural network for supervised learning as well as a neural network for unsupervised clustering. We applied the supervised network model to the credit card promotion data to test the aforementioned hypothesis. Once again, *life insurance promotion* was designated as the output attribute. Because we wanted to construct a predictive model, the input attributes were limited to *income range*, *credit card insurance*, *sex*, and *age*. Therefore the network architecture contained four input nodes and one output node. For our experiment we chose five hidden-layer nodes. Because neural networks cannot accept categorical data, we transformed categorical attribute values by replacing *yes* and *no* with 1 and 0 respectively, *male* and *female* with 1 and 0, and income range values with the lower end of each range score.

Computed and actual (desired) values for the output attribute *life insurance promotion* are shown in Table 2.4. Notice that in most cases, a computed output value is within .03 of the actual value. To use the trained network to classify a new unknown instance, the attribute values for the unknown instance are passed through the network and an output score is obtained. If the computed output value is closer to 0, we predict the instance to be an unlikely candidate for the life insurance promotion. A value closer to 1 shows the unknown instance as a good candidate for accepting the life insurance promotion.

A major shortcoming of the neural network approach is a lack of explanation about what has been learned. Converting categorical data to numerical values can also be a challenge. Chapter 8 details two common neural network learning techniques. In Chapter 9 you will learn how to use your iDA neural network software package.

## Statistical Regression

**Statistical regression** is a supervised learning technique that generalizes a set of numeric data by creating a mathematical equation relating one or more input attributes to a single numeric output attribute. A **linear regression** model is characterized by an output attribute whose value is determined by a linear sum of weighted input attribute values. Here is a linear regression equation for the data in Table 2.3:

$$\textit{life insurance promotion} = 0.5909 \leftrightarrow (\textit{credit card insurance}) - 0.5455 \leftrightarrow (\textit{sex}) + 0.7727$$

Notice that *life insurance promotion* is the attribute whose value is to be determined by a linear combination of attributes *credit card insurance* and *sex*. As with the neural network model, we transformed all categorical data by replacing *yes* and *no* with 1 and 0, *male* and *female* with 1 and 0, and income range values with the lower end of each range score.

Table 2.4 • **Neural Network Training: Actual and Computed Output**

Instance Number	Life Insurance Promotion	Computed Output
1	0	0.024
2	1	0.998
3	0	0.023
4	1	0.986
5	1	0.999
6	0	0.050
7	1	0.999
8	0	0.262
9	0	0.060
10	1	0.997
11	1	0.999
12	1	0.776
13	1	0.999
14	0	0.023
15	1	0.999

To illustrate the use of the equation, suppose we wish to determine if a female who does not have credit card insurance is a likely candidate for the life insurance promotion. Using the equation, we have:

$$\begin{aligned} \text{life insurance promotion} &= 0.5909(0) - 0.5455(0) + 0.7727 \\ &= 0.7727 \end{aligned}$$

Because the value 0.7727 is close to 1.0, we conclude that the individual is likely to take advantage of the promotional offer.

Although regression can be nonlinear, the most popular use of regression is for linear modeling. Linear regression is appropriate provided the data can be accurately modeled with a straight line function. Excel has built-in functions for performing several statistical operations, including linear regression. In Chapter 10 we will show you how to use Excel's LINEST function to create linear regression models.

## 2.3 Association Rules

---

As the name implies, **association rule** mining techniques are used to discover interesting associations between attributes contained in a database. Unlike traditional production rules, association rules can have one or several output attributes. Also, an output attribute for one rule can be an input attribute for another rule. Association rules are a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. For this reason a limited number of attributes are able to generate hundreds of association rules.

We applied the *apriori* association rule algorithm described by Agrawal et al. (1993) to the data in Table 2.3. The algorithm examines baskets of items and generates rules for those baskets containing a minimum number of items. The *apriori* algorithm does not process numerical data. Therefore, before application of the algorithm, we transformed the attribute *age* to the set of discrete categories: *over15*, *over20*, *over30*, *over40*, and *over50*. To illustrate, an individual with *age* = *over40* is between the ages of 40 and 49 inclusive. Once again, we limited the choice of attributes to *income range*, *credit card insurance*, *sex*, and *age*. Here is a list of three association rules generated by the *apriori* algorithm for the data in Table 2.3.

1. **IF** *Sex* = *Female* & *Age* = *over40* & *Credit Card Insurance* = *No*  
**THEN** *Life Insurance Promotion* = *Yes*
2. **IF** *Sex* = *Male* & *Age* = *over40* & *Credit Card Insurance* = *No*  
**THEN** *Life Insurance Promotion* = *No*
3. **IF** *Sex* = *Female* & *Age* = *over40*  
**THEN** *Credit Card Insurance* = *No* & *Life Insurance Promotion* = *Yes*

Each of these three rules has an accuracy of 100% and covers exactly 20% of all data instances. For rule 3, the 20% rule coverage tells us that one in every five individuals is a female over the age of 40 who does not have credit card insurance and has life insurance obtained through the life insurance promotional offer. Notice that in rule 3 *credit card insurance* and *life insurance promotion* are both output attributes. As the values for *age* were modified, it is difficult to compare and contrast the rules presented here to the previous rules generated by the iDA rule generator.

A problem with association rules is that along with potentially interesting rules, we are likely to see several rules of little value. In Chapter 3 we will explore this issue in more detail and describe the *apriori* algorithm for generating association rules. The next section continues our discussion by exploring unsupervised clustering techniques.

## 2.4 Clustering Techniques

---

Several unsupervised clustering techniques can be identified. One common technique is to apply some measure of similarity to divide instances into disjoint partitions. The partitions are generalized by computing a group mean for each cluster or by listing a most typical subset of instances from each cluster. In Chapter 3 we will examine an unsupervised algorithm that partitions data in this way. A second approach is to partition data in a hierarchical fashion where each level of the hierarchy is a generalization of the data at some level of abstraction. One of the unsupervised clustering models that comes with your iDA software tool is a hierarchical clustering system.

We applied the iDA unsupervised clustering model to the data in Table 2.3. Our choice for input attributes was again limited to *income range*, *credit card insurance*, *sex*, and *age*. We set the *life insurance promotion* attribute to “display only,” meaning that although the attribute is not used by the clustering system, it will appear as part of the summary statistics. As learning is unsupervised, our hypothesis needs to change. Here is a possible hypothesis that is consistent with our theme of determining likely candidates for the life insurance promotion:

*By applying unsupervised clustering to the instances of the Acme Credit Card Company database, we will find a subset of input attributes that differentiate cardholders who have taken advantage of the life insurance promotion from those cardholders who have not accepted the promotional offer.*

As you can see, we are using unsupervised clustering to find a best set of input attributes for differentiating current customers who have taken advantage of the special promotion from those who have not. Once we determine a best set of input attributes, we can use the attributes to develop a supervised model for predicting future outcomes.

To test the hypothesis, we apply unsupervised clustering to the data several times until we find a set of input attributes that result in clusters which differentiate the two classes. The results of one such clustering are displayed in Fig. 2.3. The figure indicates that three clusters were formed. As you can see, the three individuals represented in cluster 1 did not take advantage of the life insurance promotion. Two of the individuals in cluster 2 took advantage of the promotion and three did not. Finally, all seven individuals in cluster 3 purchased the life insurance promotion. Here is a production rule generated by RuleMaker for the third cluster shown in Fig. 2.3:

Figure 2.3 • An unsupervised clustering of the credit card database

**Cluster 1**

**# Instances:** 3  
**Sex:** Male => 3  
Female => 0  
**Age:** 43.3  
**Credit Card Insurance:** Yes => 0  
No => 3  
**Life Insurance Promotion:** Yes => 0  
No => 3

**Cluster 2**

**# Instances:** 5  
**Sex:** Male => 3  
Female => 2  
**Age:** 37.0  
**Credit Card Insurance:** Yes => 1  
No => 4  
**Life Insurance Promotion:** Yes => 2  
No => 3

**Cluster 3**

**# Instances:** 7  
**Sex:** Male => 2  
Female => 5  
**Age:** 39.9  
**Credit Card Insurance:** Yes => 2  
No => 5  
**Life Insurance Promotion:** Yes => 7  
No => 0

**IF** *Sex = Female & 43 >= Age >= 35 & Credit Card Insurance = No*  
**THEN** *Class = 3*  
Rule Accuracy: 100.00%  
Rule Coverage: 66.67%

It is clear that two of the three clusters differentiate individuals who took advantage of the promotion from those who did not. This result offers positive evidence that the attributes used for the clustering are viable choices for building a predictive supervised learner model. In Chapter 4 we will detail unsupervised hierarchical clustering when we investigate the ESX data mining model. In the next section we lay the foundation for evaluating the performance of supervised and unsupervised learner models.

## 2.5 Evaluating Performance

---

Performance evaluation is probably the most critical of all the steps in the data mining process. In this section we offer a common sense approach to evaluating supervised and unsupervised learner models. In later chapters we will concentrate on more formal evaluation techniques. As a starting point, we pose three general questions:

1. Will the benefits received from a data mining project more than offset the cost of the data mining process?
2. How do we interpret the results of a data mining session?
3. Can we use the results of a data mining process with confidence?

All three questions are difficult to answer. However, the first is more of a challenge because several factors come into play. Here is a minimal list of considerations for the first question:

1. Is there knowledge about projects similar to the proposed project? What are the success rates and costs of projects similar to the planned project?
2. What is the current form of the data to be analyzed? Does the data exist or will it have to be collected? When a wealth of data exists and is not in a form amenable for data mining, the greatest project cost will fall under the category of data preparation. In fact, a larger question may be whether to develop a data warehouse for future data mining projects.
3. Who will be responsible for the data mining project? How many current employees will be involved? Will outside consultants be hired?

4. Is the necessary software currently available? If not, will the software be purchased or developed? If purchased or developed, how will the software be integrated into the current system?

As you can see, any answer to the first question requires knowledge about the business model, the current state of available data, and current resources. Therefore we will turn our attention to providing evaluation tools for questions 2 and 3. We first consider the evaluation of supervised learner models.

## Evaluating Supervised Learner Models

Supervised learner models are designed to classify, estimate, and/or predict future outcome. For some applications the desire is to build models showing consistently high predictive accuracy. The following three applications focus on classification correctness:

- Develop a model to accept or reject credit card applicants
- Develop a model to accept or reject home mortgage applicants
- Develop a model to decide whether or not to drill for oil

Classification correctness is best calculated by presenting previously unseen data in the form of a test set to the model being evaluated. Test set model accuracy can be summarized in a table known as a **confusion matrix**. To illustrate, let's suppose we have three possible classes:  $C_1$ ,  $C_2$ , and  $C_3$ . A generic confusion matrix for the three-class case is shown in Table 2.5.

Values along the main diagonal give the total number of correct classifications for each class. For example, a value of 15 for  $C_{11}$  means that 15 class  $C_1$  test set instances were correctly classified. Values other than those on the main diagonal represent classification errors. To illustrate, suppose  $C_{12}$  has the value 4. This means that four class  $C_1$

Table 2.5 • **A Three-Class Confusion Matrix**

### Computed Decision

	$C_1$	$C_2$	$C_3$
$C_1$	$C_{11}$	$C_{12}$	$C_{13}$
$C_2$	$C_{21}$	$C_{22}$	$C_{23}$
$C_3$	$C_{31}$	$C_{32}$	$C_{33}$

instances were incorrectly classified as belonging to class  $C_2$ . The following three rules may be helpful in analyzing the information in a confusion matrix:

- **Rule 1.** Values along the main diagonal represent correct classifications. For the matrix in Table 2.5, the value  $C_{11}$  represents the total number of class  $C_1$  instances correctly classified by the model. A similar statement can be made for the values  $C_{22}$  and  $C_{33}$ .
- **Rule 2.** Values in row  $C_i$  represent those instances that belong to class  $C_i$ . For example, with  $i = 2$ , the instances associated with cells  $C_{21}$ ,  $C_{22}$ , and  $C_{23}$  are all actually members of  $C_2$ . To find the total number of  $C_2$  instances incorrectly classified as members of another class, we compute the sum of  $C_{21}$  and  $C_{23}$ .
- **Rule 3.** Values found in column  $C_i$  indicate those instances that have been classified as members of  $C_i$ . With  $i = 2$ , the instances associated with cells  $C_{12}$ ,  $C_{22}$ , and  $C_{32}$  have been classified as members of class  $C_2$ . To find the total number of instances incorrectly classified as members of class  $C_2$ , we compute the sum of  $C_{12}$  and  $C_{32}$ .

The three applications listed at the beginning of this section represent two-class problems. For example, a credit card application is either accepted or rejected. We can use a simple two-class confusion matrix to help us analyze each of these applications.

## Two-Class Error Analysis

Consider the confusion matrix displayed in Table 2.6. Cells showing *True Accept* and *True Reject* represent correctly classified test set instances. For the first and second applications presented in the previous section, the cell with *False Accept* denotes accepted applicants that should have been rejected. The cell with *False Reject* designates rejected applicants that should have been accepted. A similar analogy can be made for the third application. Let's use the confusion matrices shown in Table 2.7 to examine the first application in more detail.

Assume the confusion matrices shown in Table 2.7 represent the test set error rates of two supervised learner models built for the credit card application problem. The confusion matrices show that each model displays an error rate of 10%. As the error rates are identical, which model is better? To answer the question we must compare the average cost of credit card payment default to the average potential loss in profit realized by rejecting individuals who are good approval candidates. Given that credit card purchases are unsecured, the cost of accepting credit card customers likely to default is more of a concern. In this case we should choose Model B because the confusion matrices tell us that this model is less likely to erroneously offer a credit

Table 2.6 • **A Simple Confusion Matrix**

	<b>Computed Accept</b>	<b>Computed Reject</b>
Accept	True Accept	False Reject
Reject	False Accept	True Reject

card to an individual likely to default. Does the same reasoning apply for the home mortgage application? How about the application where the question is whether to drill for oil? As you can see, although test set error rate is a useful measure for model evaluation, other factors such as costs incurred for false inclusion as well as losses resulting from false omission must be considered.

## Evaluating Numeric Output

A confusion matrix is of little use for evaluating supervised learner models offering numeric output. In addition, the concept of classification correctness takes on a new meaning with numeric output models because instances cannot be directly categorized into one of several possible output classes. However, several useful measures of model accuracy have been defined for supervised models having numeric output. The most common numeric accuracy measures are mean absolute error and mean squared error.

The **mean absolute error** for a set of test data is computed by finding the average absolute difference between computed and desired outcome values. In a similar manner, the **mean squared error** is simply the average squared difference

Table 2.7 • **Two Confusion Matrices Each Showing a 10% Error Rate**

<b>Model A</b>	<b>Computed Accept</b>	<b>Computed Reject</b>	<b>Model B</b>	<b>Computed Accept</b>	<b>Computed Reject</b>
Accept	600	25	Accept	600	75
Reject	75	300	Reject	25	300

between computed and desired outcome. It is obvious that for a best test set accuracy we wish to obtain the smallest possible value for each measure. Finally, the **root mean squared error (rms)** is simply the square root of a mean squared error value. *Rms* is frequently used as a measure of test set accuracy with feed-forward neural networks.

## Comparing Models by Measuring Lift

Marketing applications that focus on response rates from mass mailings are less concerned with test set classification error and more interested in building models able to extract bias samples from large populations. The hope is to select samples that will show higher response rates than the rates seen within the general population. Supervised learner models designed for extracting bias samples from a general population are often evaluated by a measure that comes directly from marketing known as **lift**. An example illustrates the idea.

Let's consider an expanded version of the credit card promotion database. Suppose the Acme Credit Card Company is about to launch a new promotional offer with next month's credit card statement. The company has determined that for a typical month, approximately 100,000 credit card holders show a zero balance on their credit card. The company has also determined that an average of 1% of all card holders take advantage of promotional offers included with their card billings. Based on this information, approximately 1000 of the 100,000 zero-balance card holders are likely to accept the new promotional offer. As zero-balance card holders do not require a monthly billing statement, the problem is to send a zero-balance billing to exactly those customers who will accept the new promotion.

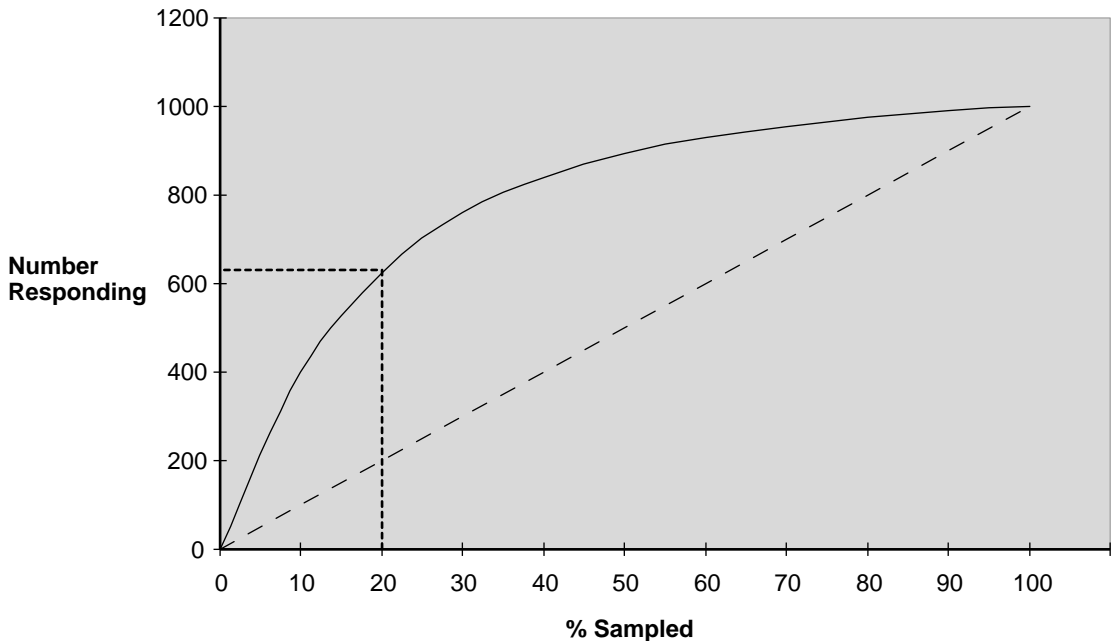
We can employ the concept of lift to help us choose a best solution. Lift measures the change in percent concentration of a desired class,  $C_i$ , taken from a biased sample relative to the concentration of  $C_i$  within the entire population. We can formulate lift using conditional probabilities. Specifically,

$$\text{Lift} = \frac{P(C_i \mid \text{Sample})}{P(C_i \mid \text{Population})}$$

where  $P(C_i \mid \text{Sample})$  is the portion of instances contained in class  $C_i$  relative to the biased sample population and  $P(C_i \mid \text{Population})$  is the fraction of class  $C_i$  instances relative to the entire population. For our problem,  $C_i$  is the class of all zero-balance customers who, given the opportunity, will take advantage of the promotional offer.

Figure 2.4 offers a graphical representation of the credit card promotion problem. The graph is sometimes called a **lift chart**. The horizontal axis shows the percent of the total population sampled and the vertical axis represents the number of

Figure 2.4 • Targeted vs. mass mailing



likely respondents. The graph displays model performance as a function of sample size. The straight line represents the general population. This line tells us that if we randomly select 20% of the population for the mailing, we can expect a response from 200 of the 1000 likely respondents. Likewise, selecting 100% of the population will give us all respondents. The curved line shows the lift achieved by employing models of varying sample sizes. By examining the graph, you can see that an ideal model will show the greatest lift with the smallest sample size. This is represented in Fig. 2.4 as the upper-left portion of the graph. Although Fig. 2.4 is useful, the confusion matrix also offers us an explanation about how lift can be incorporated to solve problems.

Table 2.8 shows two confusion matrices to help us understand the credit card promotion problem from the perspective of lift. The confusion matrix showing *No Model* tells us that all zero-balance customers are sent a billing statement with the promotional offer. By definition, the lift for this scenario is 1.0 because the sample and the population are identical. The lift for the matrix showing *Ideal Model* is 100 (100%/1%) because the biased sample contains only positive instances.

Table 2.8 • Two Confusion Matrices: No Model and an Ideal Model

No Model	Computed Accept	Computed Reject	Ideal Model	Computed Accept	Computed Reject
Accept	1,000	0	Accept	1,000	0
Reject	99,000	0	Reject	0	99,000

Table 2.9 • Two Confusion Matrices for Alternative Models with Lift Equal to 2.25

Model X	Computed Accept	Computed Reject	Model Y	Computed Accept	Computed Reject
Accept	540	460	Accept	450	550
Reject	23,460	75,540	Reject	19,550	79,450

Consider the confusion matrices for the two models shown in Table 2.9. The lift for model X is computed as:

$$Lift(\text{model X}) = \frac{540 / 2400}{1000 / 100000}$$

which evaluates to 2.25. The lift for model Y is computed as:

$$Lift(\text{model Y}) = \frac{450 / 20000}{1000 / 100000}$$

which also evaluates to 2.25. As was the case with the previous example, to answer the question about which is a better model we must have additional information about the relative costs of false negative and false positive selections. For our example, model Y is a better choice if the cost savings in mailing fees (4000 fewer mailings) more than offset the loss in profits incurred from fewer sales (90 fewer sales).

## Unsupervised Model Evaluation

Evaluating unsupervised data mining is, in general, a more difficult task than supervised evaluation. This is true because the goals of an unsupervised data mining session

are frequently not as clear as the goals for supervised learning. Here we will introduce a general technique that employs supervised learning to evaluate an unsupervised clustering and leave a more detailed discussion of unsupervised evaluation for later chapters.

All unsupervised clustering techniques compute some measure of cluster quality. A common technique is to calculate the summation of squared error differences between the instances of each cluster and their corresponding cluster center. Smaller values for sums of squared error differences indicate clusters of higher quality. However, for a detailed evaluation of unsupervised clustering, it is supervised learning that comes to the rescue. The technique is as follows:

1. Perform an unsupervised clustering. Designate each cluster as a class and assign each cluster an arbitrary name. For example, if the clustering technique outputs three clusters, the clusters could be given the class names  $C_1$ ,  $C_2$ , and  $C_3$ .
2. Choose a random sample of instances from each of the classes formed as a result of the instance clustering. Each class should be represented in the random sample in the same ratio as it is represented in the entire dataset. The percentage of total instances to sample can vary, but a good initial choice is two-thirds of all instances.
3. Build a supervised learner model using the randomly sampled instances as training data. Employ the remaining instances to test the supervised model for classification correctness.

This evaluation method has at least two advantages. First, the unsupervised clustering can be viewed as a structure supported by a supervised learner model. For example, the results of a clustering created by an unsupervised algorithm can be seen as a decision tree or a rule-based structure. A second advantage of the supervised evaluation is that test set classification correctness scores can provide additional insight into the quality of the formed clusters. We demonstrate how this technique is applied to real data in Chapters 5 and 9.

Finally, a common misconception in the business world is that data mining can be accomplished simply by choosing the right tool, turning it loose on some data, and waiting for answers to problems. This approach is doomed to failure. Machines are still machines. It is the analysis of results provided by the human element that ultimately dictates the success or failure of a data mining project. A formal KDD process model such as the one described in Chapter 5 will help provide more complete answers to the questions posed at the beginning of this section.

## 2.6 Chapter Summary

---

Data mining strategies include classification, estimation, prediction, unsupervised clustering, and market basket analysis. Classification and estimation strategies are similar in that each strategy is employed to build models able to generalize current outcome. However, the output of a classification strategy is categorical, whereas the output of an estimation strategy is numeric. A predictive strategy differs from a classification or estimation strategy in that it is used to design models for predicting future outcome rather than current behavior. Unsupervised clustering strategies are employed to discover hidden concept structures in data as well as to locate atypical data instances. The purpose of market basket analysis is to find interesting relationships among retail products. Discovered relationships can be used to design promotions, arrange shelf or catalog items, or develop cross-marketing strategies.

A data mining technique applies a data mining strategy to a set of data. Data mining techniques are defined by an algorithm and a knowledge structure. Common features that distinguish the various techniques are whether learning is supervised or unsupervised and whether their output is categorical or numeric. Familiar supervised data mining methods include decision trees, production rule generators, neural networks, and statistical methods. Association rules are a favorite technique for marketing applications. Clustering techniques employ some measure of similarity to group instances into disjoint partitions. Clustering methods are frequently used to help determine a best set of input attributes for building supervised learner models.

Performance evaluation is probably the most critical of all the steps in the data mining process. Supervised model evaluation is often performed using a training/test set scenario. Supervised models with numeric output can be evaluated by computing average absolute or average squared error differences between computed and desired outcome. Marketing applications that focus on mass mailings are interested in developing models for increasing response rates to promotions. A marketing application measures the goodness of a model by its ability to lift response rate thresholds to levels well above those achieved by naïve (mass) mailing strategies. Unsupervised models support some measure of cluster quality that can be used for evaluative purposes. Supervised learning can also be employed to evaluate the quality of the clusters formed by an unsupervised model.

## 2.7 Key Terms

---

**Association rule.** A production rule whose consequent may contain multiple conditions and attribute relationships. An output attribute in one association rule can be an input attribute in another rule.

**Classification.** A supervised learning strategy where the output attribute is categorical. Emphasis is on building models able to assign new instances to one of a set of well-defined classes.

**Confusion matrix.** A matrix used to summarize the results of a supervised classification. Entries along the main diagonal represent the total number of correct classifications. Entries other than those on the main diagonal represent classification errors.

**Data mining strategy.** An outline of an approach for problem solution.

**Data mining technique.** One or more algorithms together with an associated knowledge structure.

**Dependent variable.** A variable whose value is determined by a combination of one or more independent variables.

**Estimation.** A supervised learning strategy where the output attribute is numeric. Emphasis is on determining current rather than future outcome.

**Independent variable.** Input attributes used for building supervised or unsupervised learner models.

**Lift.** The probability of class  $C_i$  given a sample taken from population  $P$  divided by the probability of  $C_i$  given the entire population  $P$ .

**Lift chart.** A graph that displays the performance of a data mining model as a function of sample size.

**Linear regression.** A supervised learning technique that generalizes numeric data as a linear equation. The equation defines the value of an output attribute as a linear sum of weighted input attribute values.

**Market basket analysis.** A data mining strategy that attempts to find interesting relationships among retail products.

**Mean absolute error.** For a set of training or test set instances, the mean absolute error is the average absolute difference between classifier predicted output and actual output.

**Mean squared error.** For a set of training or test set instances, the mean squared error is the average of the sum of squared differences between classifier predicted output and actual output.

**Neural network.** A set of interconnected nodes designed to imitate the functioning of the human brain.

**Outliers.** Atypical data instances.

**Prediction.** A supervised learning strategy designed to determine future outcome.

**Root mean squared error.** The square root of the mean squared error.

**RuleMaker.** A supervised learner model for generating production rules from data.

**Statistical regression.** A supervised learning technique that generalizes numerical data as a mathematical equation. The equation defines the value of an output attribute as a sum of weighted input attribute values.

## 2.8 Exercises

---

### Review Questions

1. Differentiate between the following terms:
  - a. data mining technique and data mining strategy
  - b. dependent variable and independent variable
2. Can a data mining strategy be applied with more than one data mining technique? Can a data mining technique be used for more than one strategy? Explain your answers.
3. State whether each scenario is a classification, estimation, or prediction problem.
  - a. Determine a freshman's likely first-year grade point average from the student's combined Scholastic Aptitude Test (SAT) score, high school class standing, and the total number of high school science and mathematics credits.
  - b. Develop a model to determine if an individual is a good candidate for a home mortgage loan.
  - c. Create a model able to determine if a publicly traded company is likely to split its stock in the near future.
  - d. Develop a profile of an individual who has received three or more traffic violations in the past year.
  - e. Construct a model to characterize a person who frequently visits an online auction site and makes an average of at least one online purchase per month.
4. For each task listed in question 2:
  - a. Choose a best data mining technique. Explain why the technique is a good choice.
  - b. Choose one technique that would be a poor choice. Explain why the technique is a poor choice.
  - c. When appropriate, develop a list of candidate attributes.

5. Several data mining techniques were presented in this chapter. If an explanation of what has been learned is of major importance, which data mining techniques would you consider? Which of the presented techniques do not explain what they discover?
6. Suppose you have used data mining to develop two alternative models designed to accept or reject home mortgage applications. Both models show an 85% test set classification correctness. The majority of errors made by model A are *false accepts* whereas the majority of errors made by model B are *false rejects*. Which model should you choose? Justify your answer.
7. Suppose you have used data mining to develop two alternative models designed to decide whether or not to drill for oil. Both models show an 85% test set classification correctness. The majority of errors made by model A are *false accepts* whereas the majority of errors made by model B are *false rejects*. Which model should you choose? Justify your answer.
8. Explain how unsupervised clustering can be used to evaluate the likely success of a supervised learner model.
9. Explain how supervised learning can be used to help evaluate the results of an unsupervised clustering.

## Data Mining Questions

1. Draw a sketch of the feed-forward neural network applied to the credit card promotion database in the section titled Neural Networks.
2. Do you own a credit card? If so, log your card usage for the next month. Place information about each purchase in an Excel spreadsheet. Keep track of the date of purchase, the purchase amount, the city and state where the purchase was made, and a general purchase category (gasoline, groceries, clothing, etc.). In addition, keep track of any other information you believe to be important that would also be available to your credit card company. In Chapter 9 you will use a neural network to build a profile of your credit card purchasing habits. Once built, the model can be applied to new purchases to determine the likelihood that the purchases have been made by you or by someone else.

## Computational Questions

1. Consider the following three-class confusion matrix. The matrix shows the classification results of a supervised model that uses previous voting records to

determine the political party affiliation (Republican, Democrat, or Independent) of members of the United States Senate.

### Computed Decision

	Rep	Dem	Ind
Rep	42	2	1
Dem	5	40	3
Ind	0	3	4

- a. What percent of the instances were correctly classified?
  - b. According to the confusion matrix, how many Democrats are in the Senate? How many Republicans? How many Independents?
  - c. How many Republicans were classified as belonging to the Democratic Party?
  - d. How many Independents were classified as Republicans?
2. Suppose we have two classes each with 100 instances. The instances in one class contain information about individuals who currently have credit card insurance. The instances in the second class include information about individuals who have at least one credit card but are without credit card insurance. Use the following rule to answer the questions below:

**IF** *Life Insurance = Yes & Income > \$50K*

**THEN** *Credit Card Insurance = Yes*

Rule Accuracy = 80%

Rule Coverage = 40%

- a. How many individuals represented by the instances in the class of credit card insurance holders have life insurance and make more than \$50,000 per year?
  - b. How many instances representing individuals who do not have credit card insurance have life insurance and make more than \$50,000 per year?
3. Consider the confusion matrices shown on the following page.
- a. Compute the lift for Model X.
  - b. Compute the lift for Model Y.

<b>Model X</b>	<b>Computed Accept</b>	<b>Computed Reject</b>	<b>Model Y</b>	<b>Computed Accept</b>	<b>Computed Reject</b>
Accept	46	54	Accept	45	55
Reject	2,245	7,655	Reject	1,955	7,945

4. A certain mailing list consists of  $P$  names. Suppose a model has been built to determine a select group of individuals from the list who will receive a special flyer. As a second option, the flyer can be sent to all individuals on the list. Use the notation given in the confusion matrix below to show that the *lift* for choosing the model over sending the flyer to the entire population can be computed with the equation:

$$Lift = \frac{C_{11}P}{(C_{11} + C_{12})(C_{11} + C_{21})}$$

<b>Send Flyer?</b>	<b>Computed Send</b>	<b>Computed Don't Send</b>
Send	$C_{11}$	$C_{12}$
Don't Send	$C_{21}$	$C_{22}$

