ANSWERS TO CHAPTER 1 EXERCISES

Review Questions

1. Differentiate between the following terms:
   a. Data analytics is a process that builds models to solve problems. It involves acquiring, preprocessing and modeling data. Models are tested and results are reported and applied. Data mining can be a component part of a data analytics process as it is often used for modeling data.
   b. Training data is used to build a supervised learner model. Test data is applied to test the accuracy of a created model.
   c. An input attribute is used to help create a model that best differentiates the values of one or more output attributes.
   d. Shallow knowledge represents facts that can be found using simple database queries. Hidden knowledge represents structure in data that is usually not uncovered with simple database queries.
   e. The exemplar view hypothesizes that we store examples of concepts to help categorize unknown instances. The probabilistic view states that we build and store generalizations of concept instances for purposes of future categorization.
   f. The classical view requires all concept-defining properties to be present for an instance to be categorized with a specific concept. In contrast, the probabilistic view does not have rigid requirements for individual attribute values. Rather, concept definitions are stored as generalizations. An instance is categorized with the concept whose generalization most closely matches its attribute values.
   g. Supervised learning requires instances to have one or more output attributes. The purpose of a supervised learner model is to determine correct values for the output attributes. With unsupervised clustering, an output attribute is not present. The purpose of unsupervised clustering is to find a best-fit partition of the data instances.
   h. Relative to customer purchases, the actual value of a customer or client is the amount of money spent by the client. The instrinsic value of the client is the amount the client is expected to spend.

2. Choose whether supervised learning, unsupervised clustering or database query is most suitable. As appropriate, state any intital hypotheses you would like to test. List several attributes you believe to be relevant for solving the problem.
   a. Supervised learning is the likely choice where the output attribute is *returned to work.* Possible input attributes include *age*, *current pain level*, *income*, *occupation* as well as several others.
   b. Supervised learning is a reasonable choice where the output attribute is whether a given vehicle has been involved in an accident. Input attributes to consider include: vehicle type, tire type, mileage, the plant where the vehicle was manufactured, the plant where the tires where manufactured, when the vehicle or tire was manufactured (day of week), road conditions at the time of the accident etc. An argument for database query can also be made.
   c. Supervised learning where the output attribute could be categorical or numeric depending on how the settings are represented. Possible input attributes include height, weight, age, occupation, gender, and many others.

d. Supervised learning is a good choice with the output attribute designated as player name. A typical fantasy football team has a choice of playing 2 of 4 or five running backs. There are numerous input attribute possibilities including: ranking of the opposing defensive team, performance of the running back the last time the teams met (number of touchdowns, yards rushing, receiving yards), performance of the running back during the current season, etc.

e. We first assume that each product is an attribute with two possible values. For a given customer instance and product, the value is *yes* if the product was purchased by the customer and *no* if the product was not purchased. Unsupervised clustering will help determine which products are most often purchased together. We may also consider several supervised mining sessions where each product takes its turn as an output attribute. (Assocation rules are a viable choice, however they are not introduced until Chapter 2).

f. Database query is the choice.

g. Unsupervised clustering is one choice. This assumes we are not trying to determine the value of any one attribute. However, if we wish to develop a model to determine a likely favorite spectator sport based on a person's age, height, and weight, supervised learning is the best choice. For supervised learning, the output attribute is favorite spectator sport.

3. Several possibilities exist including: whether a refund is due, a change in filing status, a minor change in how the address is represented (e.g. using an abbreviation for cirlce when it was not used with prior returns), and many others.

4. Answers will vary.

5. Medical students learn surgical procedures by observing and assisting trained doctors. As individual observations lead to generalizations about how to perform specific operations, the learning is inductive.

6. As the web site contains a wealth of information about data mining, answers will vary.

7. The IEEE Conference on Data Mining identified the top 10 ML algorithms as
   Decision Trees
   K-Means (unsupervised clustering)
   Support Vector Machines (SVM)
   Apriori algorithm
   Expectation Maximization (EM)
   PageRank
   AdaBoost
   k-Nearest Neighbors (kNN)
   Naive Bayes
   Classification and Regression Tree (CART)

   There are Web links that show variations in the list.

8. There are several possibilities.
   a. Here are some choices: GPA, total number of earned credits, number of years in school, average credits taken per semester, extra curricular activities, and whether the person has a job.
   b. One classical view definition is:

      A GPA of 3.0 or higher, averages 12 or more credits per semester and has a declared major.
   c. A probabilistic definition can be stated as:

      An above average GPA, usually carries an average to above average credit load, and often times has a declared major.
   d. An exemplar view definition lists several examples and non-examples of good students. A new instance is given the classification associated with the best matching exemplar.

9. Let's pick *sore throat* as the top-level node. The only possibilities are *yes* and *no*. Instances one, three four, eight, and ten follow the *yes* path. The no path shows instances 2,5,6,7 & 9. The path for *sore throat = yes* has representatives from all three classes as does *sore throat = no*.

   Next we follow the *sore throat = yes* path and choose *headache*. We need only concern ourselves with instances 1,3,4, 8 & 10. For *headache = yes* we have instances 1 (strep throat) ,8 (allergy), & 10 (cold). For *headache = no* we have instances 3 (cold) and 4 (strep throat).

   Next follow *headache = yes* and choose *congestion*—the only remaining attribute. All three instances show *congestion = yes*, therefore the tree is unable to further differentiate the three instances. A similar problem is seen by following *headache = no*. Therefore, the path following *sore throat = yes* is unable to differentiate any of the five instances. The problem repeats itself for the path *sore throat = no*. In general, any top-level node choice of sore throat, congestion, or headache gives a similar result.

ANSWERS TO CHAPTER 2 EXERCISES

Review Questions

1. Differentiate between the terms:
   a. A data mining strategy is a template for problem solving. A data mining techique involves the application of a strategy to a specific set of data.
   b. A set if independent variables is used to build a model to determine the values of one or more dependent variables.

2. Yes on both counts. As one example, feed-forward neural networks and linear regression models can both be used for estimation problems. Likewise, neural networks can be used for estimation, classification and prediction problems. It is the nature of the data, not the data mining technique, that determines the data mining strategy.

3. Is each scenario a classification, estimation or prediction problem?
   a. This is a prediction problem as we are trying to determine future behavior.
   b. This is a classification problem as we are classifying individuals as a good or poor secured credit risks.
   c. This is a prediction problem.
   d. This is a classification problem as the violations have already occurred.
   e. This is a classification or estimation problem depending on how the output variable(s) are represented.

4. There are no absolute answers for this question. Here are some possibilities.
   a. For 3a: As there is an output attribute and the attribute is numeric, a neural network is a good choice. Statistical regression is also a possibility.

      For 3b, 3d, 3e: A decision tree model or a production rule generator is a good choice as an output attribute exists and we are likely to be interested in how the model reaches its conclusions.

      For 3c: A neural network model is a best choice as the output can be interpreted as the probability of a stock split.
   b. For 3a: Any technique limited to categorical output attributes would be of limited use as we are interested in a numeric output.

      For 3b, 3d, 3e: Any technique that does not explain its behavior is a poor choice.

      For 3c: A numeric output between 0 and 1 inclusive that can be treated as a probability of a stock split allows us to make a better determination of whether a stock is likely to split. Therefore, any technique whose output attribute must be categorical is a poor choice.
   c. Answers will vary.

5. For supervised learning decision trees, production rules and association rules provide information about the relationships seen between the input and output attributes. Neural networks and regression

models do a poor job of explaining their behavior. Various approaches to unsupervised clustering have not been discussed at this point.

6. As home mortgages represent secured credit, model A is likely to be the best choice. However, if the lender's cost for carrying out a forclosure is high, model B may be a better alternative.

7. As the cost of drilling for oil is very high, Model B is the best choice.

8. If clusters that differentiate the values of the output attribute are formed, the attributes are appropriate.

9. Each formed cluster is designated as a class. A subset of the instances from each class are used to build a supervised learner model. The remaining instances are used for testing the supervised model. The test set accuracy of the supervised model will help determine if meaningful clusters have been formed.

Data Mining Questions

1. The architecture of the network should appear similar to the network shown in Figure 2.2. However, the network should have 4 input-layer nodes, 5 hidden-layer nodes, and 1 output-layer node.

2. Students find this to be an interesting exercise. You may wish to discuss credit card billing categories and help students set up their individual spreadsheets.

Computational Questions

1. Consider the following three-class confusion matrix.
   a. 86%
   b. 48, 45, 7
   c. 2
   d. 0

2. Suppose we have two classes each with 100 instances.
   a. 40
   b. 8

3. Consider the confusion matrices shown below.
   a. 2.008
   b. 2.250

4. Let + represent the class of individuals responding positively to the flyer. Then we have,

Dividing the first fraction by the second fraction and simplifying gives the desired result.

$$P(+ \mid Sample) = \frac{C_{11}}{C_{11} + C_{21}}$$

$$P(+ \mid Population) = \frac{C_{11} + C_{12}}{P}$$

5. Any instances from the class stated in the rule consequent that satisfy the rule count toward rule accuracy as well as all instances from the competing class that do not satisfy the antecedent conditions also satisfy the rule. This second set of instances are said to satisfy the rule as they do not contradict what the rule says. The sum of these values is then divided by the total number of instances in the data set.

   Accuracy of rule 2 is 3 (number satisfying the rule antecendent with LIP = No) + 9 (number not satisfying the rule antecent with LIP = yes) = 12 /15 = 80%

   Accuracy of rule 3 is 3 (number satisfying the rule antecendent with LIP = Yes) + 6 (number not satisfying the rule antecent with LIP = No) = 9 /15 = 60%

   Accuracy of rule 4 is 4 (number satisfying the rule antecendent with LIP = No) + 8 (number not satisfying the rule antecent with LIP = yes) = 12 /15 = 80%