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%