Chapter 2

Multiple Choice Questions

(2.1)

1. Another name for an output attribute.
   a. predictive variable
   a. independent variable
   b. estimated variable
   c. dependent variable

2. Classification problems are distinguished from estimation problems in that
   a. classification problems require the output attribute to be numeric.
   b. classification problems require the output attribute to be categorical.
   c. classification problems do not allow an output attribute.
   d. classification problems are designed to predict future outcome.

3. Which statement is true about prediction problems?
   a. The output attribute must be categorical.
   b. The output attribute must be numeric.
   c. The resultant model is designed to determine future outcomes.
   d. The resultant model is designed to classify current behavior.

4. Which statement about outliers is true?
   a. Outliers should be identified and removed from a dataset.
   b. Outliers should be part of the training dataset but should not be present in the test data.
   c. Outliers should be part of the test dataset but should not be present in the training data.
   d. The nature of the problem determines how outliers are used.
   e. More than one of a,b,c or d is true.

(2.2)

5. Assume that we have a dataset containing information about 200 individuals. One hundred of these individuals have purchased life insurance. A supervised data mining session has discovered the following rule:

   IF age < 30 & credit card insurance = yes
   THEN life insurance = yes
   Rule Accuracy: 70%
   Rule Coverage: 63%

How many individuals in the class *life insurance= no* have credit card insurance and are less than 30 years old?

a. 63
b. 70
c. 30
d. 27

6. Which statement is true about neural network and linear regression models?
   a. Both models require input attributes to be numeric.
   b. Both models require numeric attributes to range between 0 and 1.
   c. The output of both models is a categorical attribute value.
   d. Both techniques build models whose output is determined by a linear sum of weighted input attribute values.
   e. More than one of a,b,c or d is true.

(2.3)

7. Unlike traditional production rules, association rules
   a. allow the same variable to be an input attribute in one rule and an output attribute in another rule.
   b. allow more than one input attribute in a single rule.
   c. require input attributes to take on numeric values.
   d. require each rule to have exactly one categorical output attribute.

(2.4)

8. Which of the following is a common use of unsupervised clustering?
   a. detect outliers
   b. determine a best set of input attributes for supervised learning
   c. evaluate the likely performance of a supervised learner model
   d. determine if meaningful relationships can be found in a dataset
   e. All of a,b,c, and d are common uses of unsupervised clustering.

(2.5)

9. The average positive difference between computed and desired outcome values.
   a. root mean squared error
   b. mean squared error
   c. mean absolute error
   d. mean positive error

10. Given desired class C and population P, lift is defined as
    a. the probability of class *C* given population *P* divided by the probability of *C* given a sample taken from the population.
    b. the probability of population *P* given a sample taken from *P*.
    c. the probability of class *C* given a sample taken from population *P*.
    d. the probability of class *C* given a sample taken from population *P* divided by the probability of *C* within the entire population *P*.

Fill in the Blank

Use the three-class confusion matrix below to answer questions 1 through 3.

	Computed Decision		
	Class 1	**Class 2**	**Class 3**
Class 1	10	5	3
Class 2	5	15	3
Class 3	2	2	5

1. What percent of the instances were correctly classified?
2. How many *class 2* instances are in the dataset?
2. How many instances were incorrectly classified with *class 3*?

Use the confusion matrix for Model X and confusion matrix for Model Y to answer questions 4 through 6.

Model X	Computed Accept	Computed Reject		**Model Y**	Computed Accept	Computed Reject
Accept	10	5		Accept	6	9
Reject	25	60		Reject	15	70

3. How many instances were classified as an accept by Model X?
4. Compute the lift for Model Y.
5. You will notice that the lift for both models is the same. Assume that the cost of a false reject is significantly higher than the cost of a false accept. Which model is the better choice?

Answers to Chapter 2 Questions

Multiple Choice Questions

1. d
2. b
3. c
4. d
5. d
6. a
7. a
8. e
9. c
10. d

Fill in the Blank

1. 60%
2. 23
3. 6
4. 35
5. 8/7
6. Model X