***Chapter 1***

1. Discuss whether or not each of the following activities is a data mining task.
   1. Dividing the customers of a company according to their gender
   2. Dividing the customers of a company according to their profitability
   3. Computing the total sales of a company
   4. Sorting a student database based on student identification numbers
   5. Predicting the outcomes of tossing a (fair) pair of dice.
   6. Predicting the future stock price of a company using historical records
   7. Monitoring the heart rate of a patient for abnormalities
   8. Monitoring seismic waves for earthquake activities
   9. Extracting the frequencies of a sound wave

***Chapter 2***

1. Classify the following attributes as binary, discrete or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

***Example:*** Age in years. ***Answer:*** Discrete, quantitative, ratio

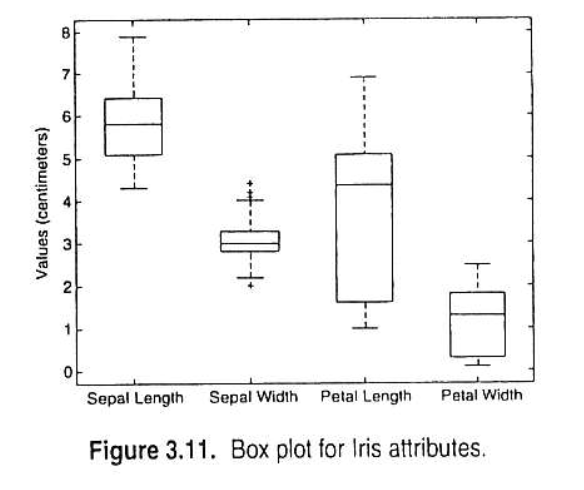
* 1. Time in terms of AM/PM
  2. Brightness measured by a light meter
  3. Brightness measured by people’s judgments
  4. Angles as measured in degrees between 0 and 360
  5. Bronze, Silver & Gold Medalists as awarded at the Olympics
  6. Height above the sea level
  7. Number of patients at a hospital
  8. ISBN Numbers for Books
  9. Ability to pass light in terms of the following values: opaque, translucent transparent
  10. Military Rank
  11. Distance from the center of Campus
  12. Density of a substance in grams per cubic centimeter
  13. Coat Check Number

1. Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

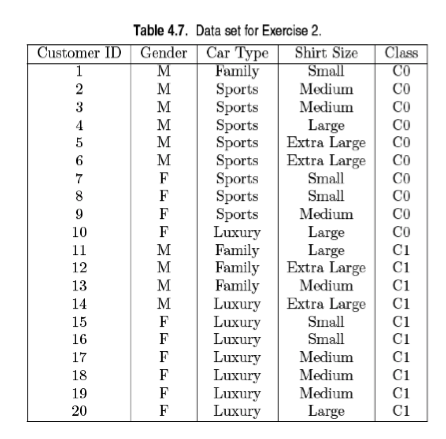
15. You are given a set of M object divided into K groups, where the ith group is of size mi. If the goal is to obtain a sample size of n<m, what is the difference between the following two sample schemes? (Assume sampling with replacement)

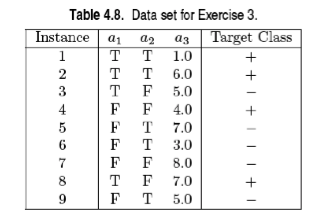
1. We randomly select n \* mi / m elements from each group
2. We randomly select n elements from the data set, without regard for the group to which an object belong
3. Assume that we apply a square root transformation to a ratio attribute x to obtain the new attribute x∗. As part of your analysis, you identify an interval (a, b) in which x∗ has a linear relationship to another attribute y.
   1. What is the corresponding interval (a, b) in terms of x?
   2. Give an equation that relates y to x.

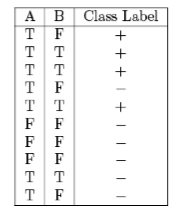
***Chapter 3***

1. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in figure 3.11?

***Chapter 4***

1. Consider the training examples shown in Table 4.7 for a binary classification problem.
   1. Compute the Gini index for the overall collection of training examples
   2. Compute the Gini index for the Customer ID attribute
   3. Compute the Gini index for the Gender attribute
   4. Compute the Gini index for the Car Type attribute using multiway split.
   5. Compute the Gini Index for the Shirt Size attribute using multiway split.
   6. Which attribute is better, Gender, Car Type, or Shirt Size?
   7. Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.
2. Consider the training examples shown in Table 4.8 for a binary classification problem.
   1. What is the entropy of this collection of training examples with respect to the positive class?
   2. What are the information gains of a1 & a2 relative to these training examples?
   3. For a3, which is a continuous attribute, compute the information gain for every possible split.
   4. What is the best split (among a1, a2, and a3) according to the information gain?
   5. What is the best split (between a1 and a2) according to the classification error rate?
   6. What is the best split (between a1 and a2) according to the Gini Index?



1. Consider the following data set for a binary class problem.
   1. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
   2. Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
   3. Figure 4.13 shows the entropy and the Gini Index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. It is possible that information gain and the gain in the Gini index favor different attributes? Explain.

