Data Mining

Problem Set 1

Chapter 1 (pg 16)

1. Discuss whether or not each of the following activities is a data mining task.

   (a) Dividing the customers of a company according to their gender.
   (b) Dividing the customers of a company according to their profitability.
   (c) Computing the total sales of a company.
   (d) Sorting a student database based on student identification numbers.
   (e) Predicting the outcomes of tossing a (fair) pair of dice.
   (f) Predicting the future stock price of a company using historical records.
   (g) Monitoring the heart rate of a patient for abnormalities.
   (h) Monitoring seismic waves for earthquake activities.
   (i) Extracting the frequencies of a sound wave.

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Chapter 2 (pg 88)

2. 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

   (a) Time in terms of AM or PM.
   (b) Brightness as measured by a light meter.
   (c) Brightness as measured by people's judgments.
   (d) Angles as measured in degrees between 0◦ and 360◦.
   (e) Bronze, Silver, and Gold medals as awarded at the Olympics.
   (f) Height above sea level.
   (g) Number of patients in a hospital.
   (h) ISBN numbers for books. (Look up the format on the Web.)
   (i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

(j) Military rank.
(k) Distance from the center of campus.
(l) Density of a substance in grams per cubic centimeter.
(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

6. An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.

   (a) How would you convert this data into a form suitable for association analysis?
   (b) In particular, what type of attributes would you have and how many of them are there?

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

11. Give at least two advantages to working with data stored in text files instead of in a binary format.

13. Consider the problem of finding the K nearest neighbors of a data object. A programmer designs Algorithm 2.1 for this task.

Algorithm 2.2 Algorithm for finding K nearest neighbors

1. **for** i = 1 to *number of data objects* **do**
2. Find the distances of the ith object to all other objects.
3. Sort these distances in decreasing order.
   (Keep track of which object is associated with each distance.)
4. **return** the objects associated with the first K distances of the sorted list
5. **end for**

   (a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.
   (b) How would you fix this problem?

19. For the following vectors, **x** and **y**, calculate the indicated similarity or distance measures.

   (a) **x** = (1, 1, 1, 1), **y** = (2, 2, 2, 2) cosine, correlation, Euclidean
   (b) **x** = (0, 1, 0, 1), **y** = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard

(c) $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean
(d) $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard
(e) $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0, -1)$ cosine, correlation

18. This exercise compares and contrasts some similarity and distance measures.

(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

$\mathbf{x}$ = 0101010001
$\mathbf{y}$ = 0100011000

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)
(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)
(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

20. Here, we further explore the cosine and correlation measures.

(a) What is the range of values that are possible for the cosine measure?
(b) If two objects have a cosine measure of 1, are they identical? Explain.
(c) What is the relationship of the cosine measure to correlation, if any? (Hint: Look at statistical measures such as mean and standard devia tion in cases where cosine and correlation are the same and different.)
(d) Figure 2.20(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between

Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

(e) Figure 2.20(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

(f) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L2 length of 1.

(g) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.

22. Discuss how you might map correlation values from the interval [−1,1] to the interval [0,1]. Note that the type of transformation that you use might depend on the application that you have in mind. Thus, consider two applications: clustering time series and predicting the behavior of one time series given another.

23. Given a similarity measure with values in the interval [0,1] describe two ways to transform this similarity value into a dissimilarity value in the interval [0,∞].

27. Show that the distance measure defined as the angle between two data vectors, **x** and **y**, satisfies the metric axioms given on page 70. Specifically, d(**x**, **y**) = arccos(cos(**x**, **y**)).