List of Exercises: Data Mining 1
November 4th, 2015

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

   (a) Time in terms of AM or PM.

   (b) Brightness as measured by a light meter.

   (c) Brightness as measured by peoples judgments.

   (d) Angles as measured in degrees between 0 and 360 degrees.

   (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.

   (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.)

2. You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and

that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints." Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?

3. 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?

4. Distinguish between noise and outliers. Be sure to consider the following questions.

   (a) Is noise ever interesting or desirable? Outliers?

   (b) Can noise objects be outliers?

   (c) Are noise objects always outliers?

   (d) Are outliers always noise objects?

   (e) Can noise make a typical value into an unusual one, or vice versa?

5. The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of similarity measure would you use to compare or group these elephants? Justify your answer and explain any special circumstances.

6. You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i . If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

   (a) We randomly select $n \times m_i/m$ elements from each group.

   (b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

7. Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

8. Describe how you would create visualizations to display information that describes the following types of systems. Be sure to address the following issues:

   - **Representation**. How will you map objects, attributes, and relationships to visual elements?
   - **Arrangement**. Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Specific examples might be the choice of viewpoint, the use of transparency, or the separation of certain groups of objects.
   - **Selection**. How will you handle a large number of attributes and data objects?

   (a) Computer networks. Be sure to include both the static aspects of the network, such as connectivity, and the dynamic aspects, such as traffic.

   (b) The distribution of specific plant and animal species around the world for a specific moment in time.

   (c) The use of computer resources, such as processor time, main memory, and disk, for a set of benchmark database programs.

   (d) The change in occupation of workers in a particular country over the last thirty years. Assume that you have yearly information about each person that also includes gender and level of education.