Fall 2018: Data Mining   Assignment 1   Due: Oct. 03, 2018

Name: RUID:

Please submit both of the question sheet and your answer before the due date.

Question 1 [20 pt]

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. Speed of a vehicle measured in mph.
2. Altitude of a region.
3. Intensity of rain as indicated using the values: no rain, intermittent rain, incessant rain.
4. Brightness as measured by a light meter.
5. Barcode number printed on each item in a supermarket.

Question 2 [20 pt]

The population for a clinical study has 500 Asian, 1000 Hispanic and 500 Native American people. What is good way of sampling this population to ensure that the distribution of various sub-populations is maintained if only 100 samples have to be chosen? Give the distribution of the various sub-populations in the final sample.

Question 3 [20 pt]

Justify your answers for the following:

1. Is the Jaccard coefficient for two binary strings (i.e., string of 0s and 1s) always greater than or equal to their cosine similarity?
2. The cosine measure can range between [-1,1]. Give an example of a type of data for which the cosine measure will always be non-negative.

Question 4 [20pt]

The similarity between two undirected graphs G_1 and G_2 that have the same n vertices can be defined using:

$$S(G_1, G_2) = \frac{\sum_i min(deg(v_i \in G_1), deg(v_i \in G_2))}{2 \times max(|G_1|, |G_2|)}, \tag{1}$$

where $deg(v \in G)$ indicates the degree of a vertex v in graph G and $|G|$ indicates the number of edges in G.

If $S(G_1, G_2) = 1$, are the two graphs equivalent? Provide an example to justify your answer.

Question 5 [20pt]

For every item i in a grocery store, a set s_i is used to represent the IDs of transactions in which i is purchased. Assume that the data set to be analyzed contains hundreds of thousands of such transactions.

1. In order to analyze the proximity between any two of these sets s_i and s_j, which measure, Jaccard or Hamming, would be more appropriate and why ?
2. In order to analyze the proximity between any two of these sets s_i and s_j for items i and j that are often brought together (example: milk, bread), which measure, Jaccard or Hamming, would be more appropriate and why ?

Extra Question [10 pt]

For the data set described below, give an example of the types of data mining questions that can be asked (one for each classification, clustering, association rule mining, and anomaly detection task) and the description of the data matrix (what are the rows and columns). If necessary, briefly explain the features that need to be constructed. Note that, depending on your data-mining question, the row and column definitions may be different.
Example data: a collection of Web pages.

DM Task: Classification of web pages
Question: What type of web page? Row: A web page. Column: Vocabulary of words that appear in a web page and a class attribute that indicates whether it is a personal home page, class web page, or company's web page.

DM Task: Clustering of web pages
Question: What are the documents with similar topics? Row: A web page. Column: Vocabulary of words that appear in a web page.

DM Task: Association rule mining
Question: What are the words that appear together frequently? For example,teaching and research appear together frequently in faculty web pages. Row: A web page. Column: Vocabulary of words that appear in a web page.

DM Task: Anomaly detection
Question: Is it a legitimate web page or a web spam? A web spam is a web page created to manipulate search engines and to deceive Web users. Row: A web page. Column: Vocabulary of words and features constructed from the hyperlinks of a web page. Examples of constructed features include the fraction of hyperlinks toURLS that reside in the same network domain or in another network domain.

(a) A clinical dataset containing various measures like temperature, blood pressure, blood glucose and heart rate for each patient during every visit, along with the diagnosis information.