CS 2756, Spring 2022
Homework 1
Due: Jan 31st, 2020, 11:59 PM, Gradescope

Instructions:

- The main submission of HW 1 must be uploaded on **Gradescope**. Your submission will be graded via Gradescope platform.
- Your PDF file should be named as "lastname_firstname_hw1". For example, if your name is Jane Doe, name your file as "doe_jane_hw1".
- Please write your name (First Name, Last Name) legibly, as it appears on Canvas. Please also include your Pitt email.
- Type out your solutions on a separate blank file (using Word, Latex, etc.), and **don't forget to convert the file to PDF** before submitting. You should not type out on the original pdf file.
- Only a subset of the questions will be graded. However, such questions are not determined a priori, therefore please do your best to answer all the questions correctly.

Question 1.

Classify the following attributes as binary, discrete, or continuous. Further, classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some of the 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) Brightness in lumens as measured by a light meter.

b) Brightness as measured by people's judgments.

c) Percentage of ones in an m-by-n binary matrix (only 0/1 entries).

d) Increase in profit of the current year over the profit obtained in the year 2000.

Question 2.

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

(i). x = (0, 1, 0, 1, 0), y = (1, 0, 1, 0, 1) cosine, correlation, Euclidean

(ii) x = (1, 1, 1, 1, 1), y = (1, 1, 1, 1, 10) Cosine, Euclidean, Correlation

Now, let’s further explore the cosine and correlation measures.

b. If two objects have a cosine measure of 1, are they identical? Explain briefly.

c. Under what conditions (if any?), is the cosine measure between two vectors the same as their correlation? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

Question 3.

(a) Suppose 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 briefly. (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.)

(b) The collection of books in a library can be represented by a vector whose length is equal to the number of distinct books available at any library and whose elements indicate the number of copies of that particular book the library owns. You want to compare the similarity of two libraries based on the collection of their books.

(i) Which similarity measure is more suited for this task between cosine and correlation. Briefly justify your answer.

(ii) What is one strength and one weakness of Euclidean distance for this task? Briefly justify your answer.

Question 4.

A team of researchers is given a large data set consisting of a time series of average monthly temperatures at a million points on the surface of the globe over 20 years. The columns correspond to months, and each row is a time series. The research team analyzes this data using correlation, L1 distance, and Euclidean distance to compare the time series and find which time series are similar to one another. Specifically, they compute the correlation/distance of each time series with every other time series. However, after visualizing some time series in the data set, the team discovers that the columns (the months) in each year of the original data are out of order, i.e., instead of (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov Dec) the order is (Jan, Jul, Feb, Aug, Mar, Sep, Apr, Oct, May, Nov, Jun, Dec). Rows are in proper order.

(a) What property of temperature time series data would make it obvious that the months were not in proper order when the time series were plotted?

(b) The team member who performed the correlation/distance calculations submits a job with the properly ordered months to recalculate the correlations/distances, but after doing so, realizes that this was unnecessary. Why?

Question 5.

You are given a set of m objects that are 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 $n * m_{\mathrm{i}} / m$ elements from each group.
(b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

Question 6.

For the following questions, answer whether the given measure is a metric. Provide **either** a brief explanation, or short proof if it is, or a counterexample if you think it is not.

a. The proximity measure between two integers x and y: $|x^2 - y^2|$

b. The proximity measure between two vectors (x_1,y_1) and (x_2,y_2): $(x_1 \times x_2)^2 + (y_1 - y_2)^2$

c. Hamming distance between two binary strings of length n.

Question 7.

Answer the following questions with True/False. Give a brief explanation for your answers.

a) It is not a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute has outliers.

b) The correlation of the vectors (1, 1, 1, 1) and (2, 2, 2, 2) is 1.

c) Cosine Similarity is better than Correlation because it ignores 0-0 matches in non-binary vector data.

d) For binary vectors, mutual information is more similar to correlation than the Jaccard measure.