CSCI 5523 Homework 1 Due: February 11, 2019

Question 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.

Ratio : absolute zero

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

1. Age as measured by whether the age in years is ≤ 30 (value = 0) or greater than 30 (value = 1).
2. The speed of a vehicle measured in mph.
3. The intensity of rain as indicated using the values: no rain, intermittent rain, incessant rain.
4. Different flavors of ice cream.

Question 2

Compute the indicated similarity measures for the following vectors, x and y:

1. x = (0, -1, 1, 2, -2), y = (0, -2, 2, 4, -4)
   Cosine, Correlation, Euclidean
2. x = (0, 1, 0, 0, 0), y = (0, 1, 0, 0, 1)
   Cosine, Jaccard, Euclidean, Correlation
3. x = (-1, -1, -1, -1), y = (1, 1, 1, 1)
   Cosine, Extended Jaccard, Euclidean, Correlation

Question 3
Give a very brief explanation in support of your answers.

1. True or False. Quantitative variables are always continuous.

2. True or False. There are cases in which Euclidean distance may not be symmetric, i.e., d(x,y) ≠ d(y,x).

3. True or False. Let a1, a2, and a3 be three attributes. If correlation(a1, a2) = 0.5 and correlation(a2, a3) = 0.5, then correlation(a1, a3) = 0.5.

4. Multiple choice. Which of the following similarity measures would be most appropriate for document data.
(a) Correlation
(b) Cosine
(c) Jaccard
(d) (a) and (b)
(e) (b) and (c)

5. True or False. The Hamming distance between two binary strings (i.e., strings of 0s and 1s), will never exceed the Euclidean distance.

Question 4

Consider a group of *n* female students and another group of *n* male students. Two *n*-dimensional vectors A and B record the heights of the two groups of students respectively. Consider the variable transformation that is defined by,

$$A = (A - \text{mean}(A)) / \text{std}(A) \quad (1)$$
$$B = (B - \text{mean}(B)) / \text{std}(B) \quad (2)$$

where std stands for the standard deviation of a vector.

1. What will be the effect of this transformation?

2. How will the correlation between A and B change after this transformation?

Question 5

A group of biologists conduct a field study to evaluate the relative occurrence of different types of birds at several locations. There data is collected in a table where
1) the rows correspond to locations,
2) the columns correspond to different species of birds, and
3) the i,jth entry is the number of birds of the jth species at the ith location.

The following table indicates the representation, but is intended only for illustration.

Location/Species	Blue Jay	Crow	…	Robin	Sparrow
Alabama	0	30		50	0
Arkansas	0	15	0	0	50
…					
Wisconsin	20	10		0	0

a. To which type of data described in Chapter 2 is this data most similar?

b. Describe the attribute type.

c. What proximity measure would you use if you wanted to find areas that were similar in terms of the percentage of birds of each species? Explain.

d. Suppose that you only care about the presence or absence of a bird species at a location. How would you change the data representation and to which type of data set in Chapter 2 would this correspond?

e. For which similarity measure and distribution of birds at a location would the similarities of two locations be the same for both data representations?

Question 6

Data reduction – sampling, dimensionality reduction, or selecting a subset of features – is necessary or useful for a wide variety of reasons, but can be problematic if information necessary to the analysis is lost in the process. The following questions explore several issues at a conceptual level.

1. Assume the property of interest is the rate at which a particular event occurs, i.e., rate = number of times a particular type of event occurs / total number of all events.
   a) If the event occurs at a rate of 0.001, i.e., 0.1% of the time, then what problems, if any, would you encounter in trying to estimate the rate from a single sample of size 100?
   b) If the event occurs at a rate of 0.50, i.e., 50% of the time, then what problems, if any, would you encounter in trying to estimate the rate from a single sample of size 100?

2. You are given a data set of 10,000,000 time series, each of which records the temperature of the Earth at a particular location on the surface of the Earth daily for 10 years. The locations are arranged in a regular grid that covers the surface of the Earth. (Details of the exact nature of the grid are unimportant. The important fact is that each point has neighbors to the left and right, up and down.) Note that temperature displays considerable autocorrelation, i.e., the temperature at a given location and time is similar to that of nearby locations and times. The size of the data needs to be reduced so that you can apply your favorite data analysis algorithm. Both aggregation and sampling could be used to reduce the amount of data.

a) If you use aggregation, would you aggregate over location or time or both?
b) How would you use the spatial and temporal autocorrelation of temperature to guide you in aggregating the data?
c) If you use sampling, would you sample over location or time or both?
d) Would you prefer aggregation or sampling or both? (You can argue any of these as long as you support your answer.)

Question 7

1. You are given a set of m objects that is divided into *K* groups, where the *ith* 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 $n * m_i / m$ elements from each group.
(b) We randomly select *n* elements from the data set, without regard for the group to which an object belongs.

2. Mapping fire activity is an important problem for supporting climate and carbon cycle studies as well as forest management. Automated approaches to address this problem use reflectance data collected from satellites orbiting the Earth to learn a classifier that can differentiate burnt areas on the ground from the unburned ones. The algorithms output a probability of fire activity for each pixel on the day the satellite imagery was taken.

   The classification algorithm needs to be trained to recognize the signatures of burned

pixels (and how they are different from non-burned ones). Typically, there is a huge imbalance in the number of fire (<0.1% of the total area) and non-fire pixels in any given region. Keeping this in mind, answer the following

(a) What would happen if one were to randomly sample some locations (pixels) for training the classifier ignoring the presence/absence of fire at the location during Sampling?

(b) How would stratified sampling help here?

PRACTICE QUESTIONS

Question 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 more ambiguity.

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

- Brightness as measured by people's judgments.
- Angles as measured in degrees between 0° and 360°.
- Density of a substance in grams per cubic centimeter.
- Ranking of movies based on the rating given by the viewers. Each viewer can rate a movie on a scale of 1 to 10 and the ranks are determined based on the sum of all ratings received for a movie.

Question 2

For each of the following vectors, x and y, calculate the indicated similarity or distance measures:

1. x = (1, 1, 0, 0, 0), y = (0, 0, 0, 1, 1) - Jaccard, Cosine, Euclidean, Correlation, Mutual Information
2. x = (0, 1, 2, 4, 5, 3), y = (5, 5, 5, 5, 5, 5) - Cosine, Euclidean, Correlation
3. x = (0, 1, 2, 4, 5, 3), y = (5, 5, 5, 5, 5, 5.0001) - Cosine, Euclidean, Correlation

Question 3

Suppose you were given temperature measurements at several locations across the country over a period of time (a time series for each city).

1. What similarity measure (defined on two time series) would you use if you want to know which cities have a similar temperature profile as your current city?
2. What similarity measure would you use if you were just interested in comparing trends in the time series and not absolute temperature differences?
3. In some studies, it might be important to pay attention to just anomalies in the data. Say we transformed the real valued temperature time series at each location into a binary time series that is 1 if the temperature is anomalous and 0 otherwise. Define a similarity measure that quantifies how similar two locations are with respect to anomalous events.

Question 4

Answer True or False, and provide a brief explanation.

1. It is a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute is constant except for small variations due to noise.
2. Equal frequency discretization is always better than equal width.

Question 5

Which of the two similarity measures, Jaccard or Correlation, would you use for the following data? Briefly explain.

1. Data set of responses to True/False test questions by a set of students. For a pair of students, the measure should capture similarity between the answers of the two students.
2. Data set of languages spoken by a set of students. For a pair of students, the measure should capture the extent to which they speak the same languages.

Question 6

Decide which of the similarity measures listed in Chapter 2 would be most appropriate for the following situations and why.

1. Suppose two of your friends are numismatists (collecting coins from different countries as a hobby). You also have coins from various countries. You want to decide which friend has the most similar collection to you. Hint: You can represent each collection as a vector of length 196 of the official independent countries of the world, where the corresponding entry denotes the number of coins collected for that country. State any assumptions you make about the coin collections.
2. Suppose you measure the precipitation level in 10 widely distributed major cities across the country for each zip code every day. Similarity is to be computed between the precipitation levels in Minnesota today and the same day of last month.
3. A nutritionist wants to measure their dissimilarity of you and your friend based on following attributes: your height (in meters), weight (in pounds), and your dietary requirements (in calories). Note that the feature set includes only continuous features.

Question 7

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

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