CSci 5523	Homework 1	Due: September 21st

Question 1

[10 points] 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

- House numbers assigned for a given street
- Your calorie intake per day
- Shape of a geometric objects commonly found in geometry classes
- Routes in rock climbing

  **NOTE: Routes in rock climbing are rated using the Yosemite Decimal System (YDS). Following is a description of the class 5 range routes in YDS (Reference: Mountaineering, Freedom of the Hills, Ed Peters):

  - 5.0 to 5.4 - There are two hand and two footholds for every move; the holds become progressively smaller as the number increases.
  - 5.5 to 5.6 - The two hand- and two footholds are there, obvious to the experienced, but not necessarily so to the beginner.
  - 5.7 -The move is missing one hand- or foothold.
  - 5.8 - The move is missing two holds of the four, or missing only one but is very strenuous.
  - 5.9 - The move has only one reasonable hold which may be for either a foot or a hand.
  - 5.10 - No hand- or footholds. The choices are to pretend a hold is there, pray a lot, or go home.
  - 5.11 - After thorough inspection you conclude this move is obviously impossible; however, occasionally someone actually accomplishes it. Since there is nothing for a handhold, grab it with both hands.
  - 5.12 - The surface is as smooth as glass and vertical. No one has really ever made this move, although a few claim they have.
  - 5.13 - This is identical to 5.12 except it is located under overhanging rock."

Question 2

[9 points] 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 number of coins collected for that country.

2. Suppose you measure the precipitation level in Minnesota for each zip code every day. Similarity is to be computed between the precipitation levels in Minnesota today and same day of last month.

3. A nutritionist wants to measure the similarity between you and your friend based on following attributes: your height (in meters), weight (in pounds), your dietary requirement (in calories), and your daily activity level (low, medium, high, extreme). Note that the feature set includes continuous and discrete features.

Question 3

[14 points] 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.

a) 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.

i) 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?

ii) 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?

b) 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.

i) If you use aggregation, would you aggregate over location or time or both?

ii) How would you use the spatial and temporal autocorrelation of temperature to guide you in aggregating the data?

iii) If you use sampling, would you sample over location or time or both?

iv) Would you perform random sampling or sample in some other way in order to best represent the data?

v) Would you prefer aggregation or sampling or both? (You can argue any of these as long as you support your answer.)

Question 4

[10 points]

a) In order to compute similarity between two documents:

i) Each document can be represented as a vector of binary features, where each feature is a word of interest in the document. In this vector a 1 indicates the presence of the word and a 0 indicates its absence.

ii) Each document can be represented as a vector of term frequencies, which represent the frequency with which each term occurs in the document. (Details of exactly how this is computed are unimportant to this problem.)

Consider the following example in which there are two documents, D1 and D2, containing four words. Using the representation described in (i) the two documents are denoted as follows:
D1 = (1, 0, 0, 1), D2 = (1, 0, 1, 0). The Jaccard similarity in this case is 0.33.
Using the representation described in (ii) the two documents are denoted as follows:
D1 = (0.5, 0, 0, 0.5), D2 = (0.5, 0, 0.5, 0). The cosine similarity in this case is 0.5.

Provide an example of a pair of documents for which Jaccard similarity will be larger than cosine similarity.

b) Both the L1 distance and correlation are widely used to compare two time series. The most appropriate measure depends on the specific problem or domain requirements. Decide the proper measure for each of the following scenarios in climate data analysis:

i) We want to compare the temperatures of two cities for 100 days with respect to their levels.

ii) We want to compare the temperatures of two cities for 100 days with respect to the trends (up and down that occur in temperature during the 100 days.

Question 5

[9 points] For the following vectors, x and y, calculate the indicated similarity or distance measures:

a) x = (0, 1, 1, 2, 2), y = (0, 2, 2, 4, 4) cosine, correlation, Euclidean
b) x = (0, 1, 0, 0, 0), y = (0, 1, 0, 0, 1) Jaccard, cosine, Euclidean, correlation
c) x = (-1, -1, -1, -1), y = (1, 1, 1, 1) cosine, correlation, Euclidean