CSci 5523 **Homework 1 - Solutions** **Due: Sept 21st, 2011**

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
  - ***Answer**: Discrete, qualitative, ordinal.*
- Your calorie intake per day
  - ***Answer:** Continuous. quantitative, ratio.*
- Shape of a geometric objects commonly found in geometry classes.
  - ***Answer:** Discrete, qualitative, nominal.*
- Routes in rock climbing
  - ***Answer**: Discrete, qualitative, ordinal. Although 5.0 to 5.4 and 5.5 to 5.6 may seem to be continuous, 5.7, 5.8, 5.9 etc. should indicate the discrete nature of the variable.*

**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.

Answer: Here, we can represent the collection of coins as a vector of length 196, where each entry of that vector represents the number of coins collected for that country. However, it will be a sparse representations where most of the entries will be zero and thus, we need to discard the zeros during the similarity measure. Any kind of asymmetric measures like Jaccard, extended Jaccard or cosine measure will work for this case. However, Jaccard works only for binary vectors and thus, will not be able to take the frequencies of coins for each country into account. So, extended Jaccard and Cosine will be the most appropriate for this case.

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.

Answer: Here, we can represent the precipitation of each day as a vector of length *n,* where *n* is the number of zip codes in Minnesota. There will be a correlation among some of the variables, i.,e, those representing nearby zip codes. So, Mahalanobis distance would be more appropriate.

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.

Answer: Here, each attribute is of different type. For example, height and weight are qualitative, and activity level in ordinal variable. So, we have to compute similarity for each attribute separately and then combine them using some weights similar as algorithm 2.1 described in the book.

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?
Answer: Assume the event occurs k times. Since there is only a single sample, the estimate of the rate is $k/100$. No matter what the value of k is, the estimate will be very inaccurate.

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?
Answer: Assume the event occurs k times. Since you have only a single sample, your estimate of the rate is $k/100$. Most of the time, k will be close to 50, so our estimate of the rate will be typically be relatively accurate, especially in comparison to the previous situation. The message is that small rates are harder to estimate than larger ones. Many intermediate level general statistics books have a more formal analysis of this result.

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?

   Answer: Either or both could be used depending on the algorithm. If the algorithm you want to use is sensitive (performs poorly or has high compute time) to high dimensionality, then aggregating over time could be useful. If the algorithm is sensitive to having lots of objects, then aggregating over space could be helpful. Additional considerations that might drive the aggregation decision are a desire to do the analysis at a coarser scale and the need to use averaging to reduce the noise level in the data.

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

   Answer: Because climate data has strong temporal and spatial autocorrelation, aggregating over either to a moderate degree would not lose too much information. For instance, aggregating days into weeks or months might still retain a fair amount of information. However, aggregating months to years, would eliminate a great deal of potentially useful detail.

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

   Answer: Again sampling over both could be useful for the same reasons as given in the answer to (ii)a.

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

   Answer: Either could be used but aggregation is probably best for many applications. The most important reason is that the autocorrelation provides a simple way of guiding the aggregation, i.e., aggregate points that are close to each other in space and time. This autocorrelation also means that a more systematic approach than random sampling should be used to ensure a representative sample. Also, comparing time series of different locations would require that the same time periods are sampled for each time series .In addition, an average might be a better representative of the original points being aggregated than any sampled time series and would have the advantage of noise reduction and potentially better representation on average. Finally, aggregation in space and time are easily and systematically accomplished by aggregating locations or time periods to a coarser unit of space or time.

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.

Answer:

Cosine can take into account the frequencies of the terms while Jaccard has to work with binary vectors. In the given example, the term frequencies are assumed to be same. If you change the frequencies of the words, you would see a different cosine similarity. The Jaccard similarity would still remain the same because the word occurances are not changed, as only the frequencies are changed. The following is one such choice of frequencies when the cosine similarity is smaller than the Jaccard similarity:
D1 = (0.25, 0, 0, 0.75), D2 = (0.25, 0, 0.75, 0)

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.
   Answer: L1 distance because the comparison is focusing on the change of level (difference of absolute value).

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.
   Answer: Correlation because the comparison is focusing on the trend (difference of value with respect to their own means)

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)

Answer:

cosine: 1

correlation: 1

Euclidean: 3.16

b) x = (0, 1, 0, 0, 0), y = (0, 1, 0, 0, 1)

Answer:

Jaccard: 0.5

cosine: 0.707

Euclidean: 1

correlation: 0.612

c) x = (-1, -1, -1, -1), y = (1, 1, 1, 1):

Answer:

cosine: -1

correlation: not defined

Euclidean: 4