CS 5525 : Data Analytics I

Homework 1

Due Date: September 13th, 2016 (4:00PM) Total: 100 Points

Problem 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.
For example: Age in years. Answer: Discrete, quantitative, ratio **(10 Points)**

(a) Angles as measured in degrees between 0° and 360°.
(b) Bronze, Silver, and Gold medals as awarded at the Olympics.
(c) Cell Phone Numbers.
(d) Different departments in the University.

Problem 2. For the given vector A= [1204, -212, 30.21, 12.0, 56], give the transformed values after using the following normalization methods. **(10 Points)**
(a) z-score normalization (b) Decimal scaling (c) Min-Max normalization [0,1]

Problem 3. Now, concatenate Vector A from the previous question (**Problem 2**) to the following vector C= [30 , 50, 10, 30, 40, -25, 95] and discretize them into three bins using
(1) Equal Width Discretization and (2) Equal Frequency Discretization **(10 Points)**

Problem 4. This problem compares and contrasts some similarity and distance measures. **(10 Points; 5 Points each)**
(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

x = 0101010001
y = 0100011000

(b) Which approach (Jaccard or Hamming distance) is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

Problem 5. For the following vectors, x and y, calculate the required similarity or distance measures:
(a) Cosine (b) Correlation (c) Euclidean (d) Jaccard **(10 Points)**

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

Problem 6. Mike completes jogging one round on a (circular) athletic track of radius 1 mile. John is waiting for him at the center of the track. Compute the minimum and maximum possible values for the following distance measures between Mike and John while Mike is jogging: Manhattan, Euclidean and Chebyshev distance. **(10 Points)**

Problem 7. Give the range (Lower bound and upper bound) for the following measures (assume n-dimensional space): (i) Euclidean Distance, (ii) Cosine Similarity (iii) Simple Matching Coefficient (iv) Hamming Distance **(10 Points)**

Problem 8. Solve the following two problems. **(10 Points; 4 + 6 Points)**
(a) Two vectors x and y have zero mean. What is the relationship of the cosine measure and correlation between them?
(b) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object vector has an L2 length (magnitude) of 1. (NOTE: your final answer should be independent of the original vectors).

Problem 9. This problem is a MATLAB exercise. **(20 Points; 4 Points each)**
(Alternatively you can also do this exercise in your favorite programming environment.)
(a) Load iris.dat file (available at the course website). Give the basic description of the data matrix, including no. of data points, no. of features, no. of classes; give some basic statistics (such as mean, median, standard deviation, min, max) for each of these features.
(c) Plot the first two features of the data. Classes must be discriminated by using different symbols. Please label the figure.
(d) Apply Principal Component Analysis (PCA) to reduce the dimension of the iris dataset to 2. Give the 2-D plot of the two new features.
(e) Similarly, apply Singular Value Decomposition (SVD) to reduce the dimension of the iris dataset to 2. Give the 2-D plot of the two new features. Is there any difference between the result obtained here and the one in (d)?