CS 5593 - Fall 2019 - Dr. Le Gruenwald
HOMEWORK 1
Assigned: 9/16/2019; Due: 9/25/2019 by 12:00 PM (NOON) to the class website on Canvas
Maximum Points: 100 points

Notes:

- **Homework answers must be typed and submitted by 12:00 PM (NOON) to the class website on Canvas.**
- **Homework is individual work; it must be done by you only, no collaboration with anyone else is allowed.**
- **Late homework will be accepted on the class website on Canvas until 11:59 PM on the date following the due date with 5% (of the maximum points) penalty. Late homework submitted after this time will not be graded.**

Problem 1 (10 points): For each of the following activities, answer if it is or if it is not a data mining task, and justify your answer.

1. Find out how a stock price will change based on the stock's past performance.
2. Find out the frequency of amino acids types of human Hemoglobin protein.
3. Identify all students who are likely to get an "A" in the Operating Systems course.
4. Recommend movies to customers based on their movie viewing pattern.
5. Identify all customers who have made more than 5 purchases at a local store.

Problem 2 (15 points): For each attribute given, classify its type as:

- discrete or continuous AND
- qualitative or quantitative AND
- nominal, ordinal, interval, or ratio

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

1. Cell phone brands
2. IQ levels
3. The states of the United States
4. The prices of laptops
5. The result of whether a person has passed a Driving Exam. The result can only be "Pass" or "Fail".

Problem 3 (25 points): Perform the following tasks:

1. Calculate the similarity/distance measure between the two vectors given below using each of the five methods: Simple Matching Coefficient, Jaccard Coefficient, Cosine, Correlation, and Hamming distance. Show your work in detail.

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

2. Assume that the following set of two-dimensional points are randomly sampled from the same multivariate Gaussian distribution: {(-2.05, 2.32), (-0.41, 5.36), (0.72, 3.62), (0.15,

3.1), (-1.3, 4.1), (-3.7, 2.8)}. Compute the Euclidean distance and the Mahalanobis distance between the first two points in the set. Show your work in detail.

3. The bank (iBank) wants to analyze the results of a survey that it has conducted to study the customers' experiences with its services. The survey has 60 questions, each of which has five possible answers (Very Dissatisfied; Somewhat Dissatisfied; Neutral; Somewhat Satisfied; Very Satisfied).

    a. How would you convert this data into a form suitable for association analysis? In particular, what types of attributes would you have and how many are there in total?
    b. Answer the same question in part (a), but with the assumption that both non-zero and zero values are equally important.

Problem 4 (50 points): Using R and the Credit Approval Data Set from the UCI dataset repository web site https://archive.ics.uci.edu/ml/datasets/Credit+Approval , perform the following tasks:

1. Write a function that estimates the missing values in the dataset as follows: for each attribute that has one or more missing values, replace every missing value in that attribute with the *mean* value if the attribute is continuous and with the *mode* value if the attribute is categorical. Then use this function to estimate all the missing values in the dataset.
2. Using the six attributes A2, A3, A8, A11, A14, and A15 of the preprocessed data, do the following:
    a. Draw a random sample *with replacement* of size N=100 from the dataset.
    b. Create a matrix of scatter plots showing the correlations between all possible pairs of attributes from the above six attributes. Using the scatter plots, comment on which pairs of attributes exhibit the highest correlation. The axes of your plots must be properly labeled.
    c. Calculate the correlations between all possible pairs formed by those six attributes.
    d. Normalize those six attributes using the Z-score normalization.
    e. Create another matrix of scatter plots showing the correlations between all possible pairs of attributes from the six *normalized* attributes.
    f. Re-calculate the correlations between all possible pairs formed by the six *normalized* attributes.
3. Using your scatter plots of Task 2 as examples, explain *in detail* whether or not the Z-score normalization affects the correlation of the data.

Notes on submission: Submit one complete PDF document that contains the answers to all the questions; for Problem 4, this complete document needs to contain the R program that you used to estimate the missing data, draw the random sample, perform normalization, and output the plots as well as the output of the runs of the program. In addition, besides this complete document, submit a separate .r text file containing your R program for Problem 4 as we will test your program for correctness. Failure to submit this file will result in a zero grade for your Problem 4.