CS 5593 – Section 001 - Fall 2020 - Dr. Le Gruenwald

HOMEWORK 1

Assigned: 9/21/2020; Due: 9/30/2020 by 2:00 PM to the class website on Canvas

Maximum Points: 100 points

Notes:

- **Homework answers must be typed and submitted by 2:00 PM 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 (16 points): For the Ambulatory Medical Care data set given below, give one specific example for each of the classification, clustering, association rule mining, and anomaly detection tasks that can be performed on the data (i.e. give an example of a data mining task that would require classification; an example of a data mining task that would require clustering, etc.). For each task, state how the data matrix should be constructed (i.e., specify the rows and columns of the matrix).

The Ambulatory Medical Care dataset contains the demographic and medical visit information for each patient (e.g., gender, age, duration of visit, physician's diagnosis, symptoms, medication, etc.).

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. Daily user traffic volume at YouTube.com (i.e., number of daily visitors who visited the Web site).
2. Air pressure of a car/bicycle tire (in psi).
3. Homeland Security Advisory System ratings - code red/orange/etc.
4. Credit card number.
5. The wealth of a nation measured in terms of gross domestic product (GDP) per capita above the world's average of $10,500.

Problem 3 (15 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 Minkowski with r =1 and r = 2. Show your work in detail.

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

2. Given the following set of two-dimensional data points generated from the same distribution: {(35, 205401), (27, -148497), (33, -69783), (24, 84958), (13, -48436), (25, 159102), (43, 270076)}, which of the following distance functions is the most appropriate one to use for computing the distance between two data points in the set: Mahalanobis, Euclidean, or Simple Matching Coefficient? Provide detailed justifications.

Problem 4 (54 points): Using R and the Auto MPG Dataset from the UCI dataset repository web site https://archive.ics.uci.edu/ml/datasets/Auto+MPG, perform the following tasks:

1. Write a function that estimates the missing values in the dataset by using the means of the available values as follows: for each continuous attribute that has one or more missing values, compute the mean of the available values in that attribute and replace every missing value in that attribute with the mean. Then use this function to estimate all the missing values of continuous attributes in the dataset.
2. Draw a random sample of size 50 *without replacement* out of all rows in the dataset, and discard the rest.
3. Write a function that performs equal frequency discretization of a continuous attribute.
4. Write a function that performs equal width discretization of a continuous attribute.
5. Do the following:
   a. Use the function implemented in Task 3 to discretize the *horsepower* attribute, and plot the original values of this attribute along with the split points resulting from the discretization function (like the plot in Figure 2.13 on Page 67 in the textbook). To do this, create a new attribute named ‘random_value’ and assign a random value for this ‘random_value’ attribute for every row in the dataset, and then plot *horsepower* vs. ‘random_value’.
   b. Repeat Task 5.a, but use the discretization function implemented in Task 4.
   c. Inspect the two plots produced in Tasks 5.a and 5.b and discuss what you have observed.
6. Create a matrix of scatter plots of the continuous attributes. Of these attributes, which ones do you think exhibit the greatest variability? Do you observe any correlation between any pair of attributes? Why?

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 samples, perform discretization of continuous attributes, create the matrix of scatter plots of the continuous attributes, and output the discretization plots, the matrix of scatter plots, as well as the output of the runs of the program. In addition, besides this complete PDF 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.