Fall 2023: CS 4435/5435 and DASE 4435 Data Mining

Homework 1, Due on September 12, 2023

Instruction: The questions marked with ***asterisk** is for graduate students only. Undergraduates are free to attempt it as a bonus point. However, the bonus is awarded only if the question is answered correctly. Also, note that poor presentation of plots will lead to point detection. A good plot would include, plot legend, title, axis labels and proper figure labelling.

A) Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression. Give one example to describe your point. Please don't use the examples in the lecture slides. (**10 points**)

B) Write a program in R, python or other programming tools to solve the following. Provide the *code as a separate attachment* and comment your code appropriately.

1. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45,46, 52, 70. (**25 points**)

Note: Don't use programming library for questions a to c below.

(a) What is the *mean, median, and mode* of the data?
(b) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?
(c) Give the *five-number summary* of the data.
(d) Show a boxplot of the data. ***

2. The data here contains the Spearman Correlation Coefficient (SCC) result score obtained by accessing the three-dimensional(3D) structures of Chromosome 1 to 23 generated by different 3D chromosome reconstruction methods–3DMax, LorDG, ShNeigh2, Chromosome3D, HSA, and ChromeBat– for a GM06990 cell line. GM06990 is a cell line derived from some human lymphoblastoid cells. SCC scores are in the range -1 to 1, the closer to 1 the better.
   Using this data (**45 points**):

(a) Calculate the mean, median, and standard deviation* of *the SCC scores.*
(b) Show a boxplot for *of the SCC data. Note: show the boxplot side-by-side in the same figure not separately.*
(c) Comment on the box plot results:
   a. Which method(s) would you regard as the best performing ones, why?
   b. Which method(s) would you regard as the least performing ones, why?
   c. Does the five-number summary plotted on the boxplot provide a relevant knowledge different from what can be obtained from the mean, median and standard deviation values only? What are they?
(d) Using a Scatter plot of any three methods' result, what is the pattern of performance you can observe from these methods' results on some chromosome(s)? Provide the scatter plots in your report as well.

Code for question 1 (**10 points**)

Code for question 2 (**10 points)**