CMPE558 Data Mining

Fall 2018

**Programming Assignment#1**

**Data Preprocessing**

**Task**

In this task you are asked to apply data preprocessing techniques to the gene expression data set, *558PA1data.txt,* available on the course web page. You should download this file and write a program (or programs) to do the following:

**Dataset**

The data set contains gene expression data for 181 genes each represented by 500 attributes. The last column of each sample represents the class for each gene.

**What to do**

I-Attribute Discretization

1. Use attributes 10, 100, 200, 300, and 450 and discretize the attribute values into 4 different groups using i) equal width approach ii) equal frequency approach
2. For each case above plot the discretized set of values using the approach described in example 2.12 in the text book.

II- Attribute Similarity

1. Choose 10 genes randomly from the dataset (specify the ones chosen in your report). For all chosen gene pairs (gi,gj), compute i) the Euclidian Distance, ii) Cosine similarity and report the most similar and dissimilar gene pairs based on your results.
2. Choose 10 attributes randomly from the dataset (specify the ones chosen in your report). For all chosen attribute pairs (ai,aj), compute state the 10 most positively correlated, most negatively correlated and most uncorrelated attribute pairs.
3. Apply min-max normalization to all attributes in the data set and repeat parts (a) and (b) above. Compare and comment on the results. Min –max normalization on an attribute *A* maps a value *v*, of *A* to *v’* in the range [*new\_minA,mew\_maxA*] by computing $v^{'}=\frac{v-min\_{A}}{max\_{A}-min\_{A}}\left(new\\_max\_{A}-new\\_min\_{A}\right)+new\\_min\_{A}$

Define the range [*new\_minA,mew\_maxA* ] as [-1,1] if the attribute *A* contains both positive and negative values, as [0,1] if it contains only positive values and [-1,0] if it contains only negative values.

**Report**

You should write a short report (3-5 pages) describing the work you have done. Your report should briefly describe problem, the data used, the methods you have used to process the data and your results. The answer to all questions above must be in your report. You need to discuss the results highlighting main findings and possible reasons and conclusions. Just stating the results is never enough. Use figures and tables where necessary. You should conclude very briefly. Your report should make reference to all resources you have accessed and/or used. In case you have used readily available software you should both cite it and also briefly explain its usage to me in your report.

**Programming Language**

You are free to use any programming language you feel comfortable with. Please don’t try to discuss with me the difficulties you may encounter during programming such as reading, writing data files, fixing bugs, codes not working etc. ☺. If you use publicly available software, again make sure it is properly cited and its usage is explained in the report.

**Grading**

This assignment is worth 15% of your CMPE558 course grade.

**Working with Teams**

* You may want to do the assignment on your own or in groups of 2.
* Members of the same team will may not receive the same grade.

**Schedule:**

* Assigned date: October 22, 2018
* Due date: November 5, 2018 at 9.30 a.m in class. There will be a 10% penalty applied to your grade for every day that it is not turned in.

**What to turn in**

* Your report.
* A CD which contains your program code(s) and an electronic version of your report. All material must be given under a folder whose name is your student id.
* All material must be handed it to me in person (in class); if you turn in your report later than the due date/time, drop it down at my pigeon box at the secretary’s room if I am not around but make sure you send me an e-mail as soon as possible telling me that you have left it there. I will assume no responsibility for work slid under my door or left in the secretary’s office without an e-mail notification that follows it.

**Academic Honesty**

As stated in the CMPE558 Course outline:

“All work submitted must be of your own. You are allowed to discuss solutions to programming assignments with your friends but you must not share codes or ideas in detail. In any case you must acknowledge the person you have shared ideas with in name and in writing. For programming assignments you are normally permitted to use code that is publicly available but it must be properly acknowledged and referenced. For all written work, you must quote (using quotation marks “ ”) any sentences taken from other published material For ideas used from other written sources, the source must be properly referenced. Plagiarism is a type of cheating (if you are not sure what is considered as plagiarism, please ask) and it will be dealt accordingly. It will result in a final course grade of “F” and may be referred to the EMU disciplinary committee.”