Project 3

Submit Assignment

Due May 5 by 8am **Points** 140 **Submitting** a file upload **File Types** zip

Data Clustering

Work in groups of 3 … WILL TAKE A LOT OF TIME

Introduction

In this project, you are asked to implement your own *K*-means clustering algorithm in Python. Your program should take as input any ***.csv*** data file containing ONLY numeric attributes and a class label (will always be the very last attribute in the file BUT does not necessarily have to be limited to two values; it is NOT reasonable to expect users to manually input class label values or their number to the program).

The program should then produce multiple clusterings using different combinations of distance measures and values for *K* in an attempt to find an "optimal clustering". Use standard stopping conditions for *K*-means (i.e., until it converges) provided it does not exceed a fixed number of iterations (say 100).

A single program execution should produce a different clustering for EACH of the six possible combinations below:

> **Two distance measures**: *Euclidean* and *Manhattan* distances–exclude class labels from distance computations (again, note that class label will always be the very last attribute in the data file); class labels will only be used to do *supervised validation* as explained later on
>
> **Three values for *k***: equal to the number of possible class labels in the file, twice the number of class labels, and three times that number

Program Output

For every combination of the above, your program should:

> (1) Produce and output the *K* clusters in a meaningful way (I want to be able to see the clusters in a nice and concise manner … be creative)

(2) *Perform unsupervised validation for each clustering*: Report the cohesion and separation using *WSS* and *BSS* measures

(3) *Perform supervised validation for each clustering*: Report the Information gain for dividing the original dataset into clusters; in essence, this is similar to computing the information gain for a new split in a decision tree by viewing each resulting cluster as a node/branch

Report

Run your program 10 times on EACH of the two datasets–*AllGenes.csv* and *SigGenes.csv* datasets which can be **downloaded here**–and compute the average values for the following validity measures: *WSS*, *BSS* and *information gain* over the 10 runs.

In your report, create meaningful tables to show the effect of the value of *K* and the distance measure on the quality of the generated clusters as defined by the average *WSS*, the ratio of (average *BSS*/average *WSS*), and average *information gain* (i.e., three tables in total for each dataset). Plot your tables in meaningful graphs as you see fit.

Repeat the above for the two datasets and then use your tables and graphs to draw meaningful conclusions for the following:

(1) Do you notice the following validity measures (*WSS*, *BSS*/*WSS*, and *info gain*) agreeing for *AllGenes.csv*? Please elaborate.

(2) What about for *SigGenes.csv*? Again, please elaborate.

(3) What is the best distance measure/*K* combination according to *WSS* alone, *BSS*/*WSS*, and *information gain* for *AllGenes.csv*? Why?

(4) What about for *SigGenes.csv* and why?

(5) Overall, which of the two datasets clusters better and when?

Advanced Clustering Visualization using Tableau

Tableau is one of the most popular visualization software currently in data science. It is available on our windows boxes in all computer rooms in Main as well as the Quad computer lab. You can download a 1-year free version by signing up at **https://www.tableau.com/academic/students (https://www.tableau.com/academic/students)**

(1) Use the Tableau software to visualize the characteristics for the produced clusters (i.e., WSS, BSS, and *information gain*) in a meaningful way as a function of the distance measure chosen and value for *k*. Include snapshots of your visualizations in your report. BE CREATIVE.

(2) Include a mechanism to interactively update program parameters and observe (1) updating automatically. BE CREATIVE. You will need to research how to run Python code from within Tableau.

Submission

EACH MEMBER should make their own submission by the noted due date and time (replica of each other). Your submission will be in the form of a single zip file containing the following:

(1) Report (in PDF format): In addition to the items requested before, your report MUST include a 1-page description detailing what EACH MEMBER did along with a percentage value describing the member's share of the work (in an ideal world, in a group of 4, each member gets 25%). If you disagree with your percentage, I ask that you inform me ASAP.

(2) Complete program including ALL NECESSARY FILES for me to run your work. Also include a READ ME text file with instructions for me to run your program. Your program should work for ANY CSV DATASET containing numeric/real data (NOT NECESSARILY INTEGERS) along with a class label (**DO NOT ASSUME A TWO-CLASS PROBLEM**). During your demo, you will be asked to run your program on the given data files as well as ones that you haven't seen.