CS 5593 – Section 001 - Fall 2023 - Dr. Le Gruenwald
HOMEWORK 4
Assigned: 11/1/2023; Due: 11/13/2023 by 3:00 PM to the class website on Canvas
Maximum Points: 100 points

Notes

- **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.**
- **ACADEMIC INTEGRITY: Homework is individual work; it must be done by you only; no collaboration with anyone is allowed. The use of generative AI tools (including ChatGPT, Bard, Bing Chat, and other AI writing and coding assistants) is not allowed in all homework questions, except for those questions where the use of these tools is explicitly required. Violations of any of these rules will be considered academic misconduct and will result in action as specified in the Academic Integrity Code at The University of Oklahoma: http://www.ou.edu/integrity. Consult also the following web page for a Student's Guide to Academic Integrity at The University of Oklahoma: http://www.ou.edu/integrity/students.**

Problem 1 (15 points): Using ChatGPT, generate prompts requesting information about three real-world companies/organizations from three distinct sectors—healthcare, finance, and manufacturing—that have implemented data clustering algorithms in their businesses. Ask for detailed descriptions of how those companies/organizations have applied data clustering algorithms in their operations. You can try at most three prompts. For each real-world application description provided in ChatGPT's final answer, elaborate on whether the description is sufficient or insufficient to understand how data clustering is utilized, including the concept of a similarity measure. For each description that you think is insufficient, modify it to ensure a comprehensive understanding of how data clustering is employed. Show your prompts, ChatGPT's answers, your explanations, and your modifications (if any).

Problem 2 (15 points) (your answers to this problem can be hand-written, but must be readable in the submitted PDF version): Given the following dataset where each data point has two dimensions, X and Y, apply Bisecting K-means to cluster the data of the given dataset with K = 3 using the number of trials = 2, and choosing the cluster with the largest SSE for splitting at each step. You need to show your work manually step by step.

X	Y
6.57	19.09
8.76	22.85
1.88	3.76
7.20	20.35
5.01	31.61
5.63	-17.22
4.38	-15.96
0.69	2.50

Problem 3 (70 points): Using a programming language of your choice (e.g. Java, Python, R, etc.), perform the following tasks:

3.1) Download the Dow Jones Index dataset from https://archive.ics.uci.edu/ml/datasets/dow+jones+index.

3.2) Remove the three categorical attributes, "quarter", "stock", and "date" from the dataset, and perform any other pre-processing steps that you consider necessary. Then write the justifications for the need to apply those pre-processing steps. In case you consider that no pre-processing is needed, you must justify such decision.

3.3) Using the K-means algorithm available in the programing language of your choice (you do not need to implement this algorithm from scratch yourself; you can use the existing packages/functions that already implement this algorithm in your program if they are available in the programming language of your choice), cluster the instances of the dataset with different cluster numbers K = 2, 3, 4, 5, 6, and 7. For each value of K that you run K-means with, print out the total sum of squared errors, the sum of squared errors for each cluster, the cluster mean, the cluster ID, and the IDs of all the instances belonging to that cluster. Then, using the elbow/knee method, make a plot of the total sum of squared errors vs. K and select an adequate K value. Justify why you choose that K value.

3.4) Using the agglomerative hierarchical clustering Complete Link (MAX) algorithm and the agglomerative hierarchical clustering Centroid algorithm available in the programming language of your choice (you do not need to implement these algorithms from scratch yourself; you can use the existing packages/functions that already implement these algorithms in your program if they are available in the programming language of your choice), cluster the instances of the dataset using the Complete Link (MAX) algorithm and the Centroid algorithm. Then for each of the algorithms, using the K value you chose in Task 3.3, get the K clusters by cutting the resulting dendrogram at the K level (the root of a dendrogram is at level 1). For each clustering, print out the cut dendrogram using either a textual or a graphical format, and the total sum of squared errors, and for each cluster in the clustering, print out the cluster ID and the IDs of all the instances belonging to that cluster, and the cluster's sum of squared errors. Once you have done this, justify which inter-cluster similarity (MAX or Centroid) is the best for the task.

3.5) Write a report that presents an in-depth comparison analysis based on the results you got from Tasks 3.3 and 3.4, describing advantages and disadvantages of the K-means and the agglomerative hierarchical clustering algorithm with different similarity measures (MAX and Centroid) for classifying the given dataset. The analysis must be comprehensive, not a single sentence; it is not only to report the results, but also to explain why you get such results (e.g. the statement, *"K- mean with K = X is the best for this dataset because it achieves the smallest total sum of squared errors,"* is NOT a complete analysis; you need to explain why this occurs).

Notes on choosing a programming language to write your program for Problem 3: since Tasks 3.3 and 3.4 together require you to run three specific clustering algorithms, you should choose a programming language (e.g. R), in which there exist packages/functions that already implement those clustering algorithms so that your program only needs to use/call those packages/functions; otherwise, you will have to implement those algorithms from scratch yourself.

Notes on submission: Submit two files to the class website. The first file is a complete PDF file named *Your Last Name_Your First Name–HW4–CS5593.pdf* containing your answers to ALL the three problems. For Problem 3, this complete PDF file needs to also contain the source code of the program that you wrote to accomplish all the tasks, the screenshots of your program executions and the required output. Appropriate in-line comments must be included in the program, and

appropriate labels must be provided for the required output. The second file is a program file named *Your Last Name_Your First Name–HW4–CS5593.extension* containing the program for Problem 3 as we will test your program for correctness, where *extension* is the programming language you used to implement your program (e.g. r for a program written in R or py for a program written in Python). Failure to submit this program file will result in a zero grade for your Problem 3.