CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2024: Homework 8 (10 points)

Due date: Sunday, December 01 2024 11:59:59 PM Chicago Time

NOTE: No late exception for HW 7 is allowed. This is the last homework and it needs to be scored in time to submit the final grade to the registrar.

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (3 points divided evenly among the questions) Please submit a PDF file containing answers to these questions. Any other file format that can be read will lead to a loss of 0.5 point. Non-PDF files that cannot be opened and read for grading will lead to a loss of all points allocated to this exercise.

1.1 Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms)

Exercise 2, 6, 7, 11, 12, 16. (For 16, note that Table 7.13 for Exercise 16 has a similarity matrix, not a distance matrix. Similarity and distance are related to each other by the formula *distance = 1.0 – similarity*.)

2. Practicum problems

2.1 Hierarchical clustering

HARTIGAN is a dataset directory that contains test data for clustering algorithms. The data files are all simple text files, and the format of the data files is explained on the web page at https://people.sc.fsu.edu/~jburkardt/datasets/hartigan/hartigan.html

Perform hierarchical clustering on file46.txt on the above web page. This file 16 rows and 13 columns; each row is a country and each column is a language spoken by the percentage of people in the country. (FI – Finnish, SW – Swedish, DA – Danish, NO – Norwegian, EN – English, GE – German, DU – Dutch, FL – Flanders, FR – French, IT – Italian, SP – Spanish, PO – Portuguese.) So for example, the first row of data indicates that in West Germany 21% of the people speak English, 100% speak German, 2% speak Dutch, 1% speak Flanders, 10% speak French, 2% speak Italian, and 1% speak Spanish.

Before using the dataset, standardize it (mean = 0.0, sd = 1.0).

(a) Data cleanup

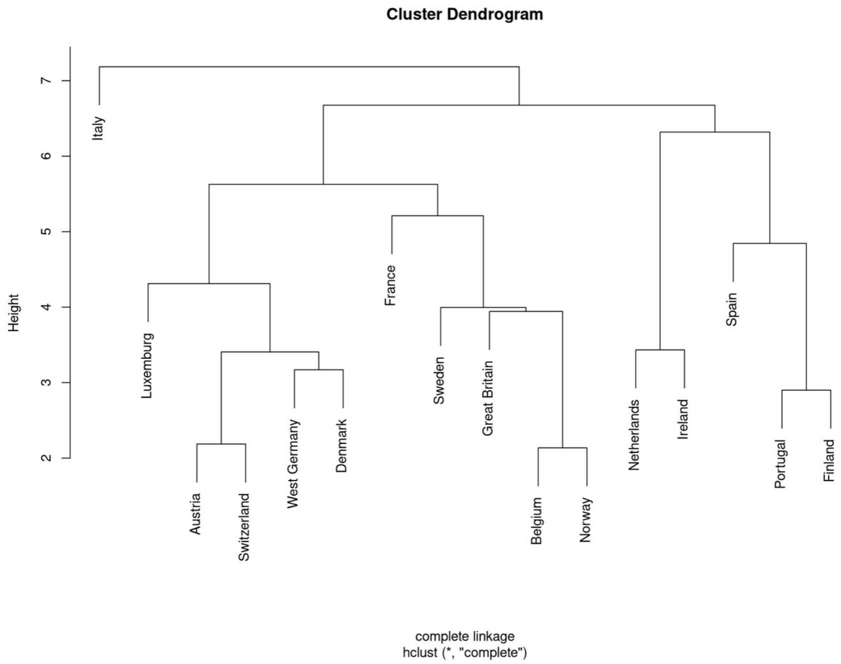
[0.5 points] Write a function to clean the data to remove multiple spaces and make the comma character the delimiter. Your function should return a data frame that has 16 rows and 13 columns, except that the values in column 2-13 are in standardized form. Please make sure you include your cleaned dataset in the archive file you upload.

(b) Hierarchical Clustering

(i) **[0.375 points]** Write a function to perform hierarchical clustering using Complete linkage. Plot the clusters using the plot() API; make sure you include the labels (the names of the countries) in the graph. The value of the x-label should be "complete linkage" (lower-cased). Do not change the value of the default y-label.

Your function should return an un-named list whose first element is the object returned to you by hierarchical clustering, and the second object is the name of your plot file.

Here is an example of the plot you should create; the image below is a plot for Complete linkage on a perturbed dataset (i.e., I randomly modified the values of the data frame; your results on complete linkage will be different than what I show below. The image below is just for illustration.)



(ii) **[0.375 points]** Write a function to perform hierarchical clustering using Single linkage. Plot the clusters using the plot() API; make sure you include the labels (the names of the countries) in the graph. The value of the x-label should be "single linkage" (lower-cased). Do not change the value of the default y-label.

Your function should return an un-named list whose first element is the object returned to you by hierarchical clustering, and the second object is the name of your plot file.

(iii) **[0.375 points]** Write a function to perform hierarchical clustering using Average linkage. Plot the clusters using the plot() API; make sure you include the labels (the names of the countries) in the graph. The value of the x-label should be "average linkage" (lower-cased). Do not change the value of the default y-label.

Your function should return an un-named list whose first element is the object returned to you by hierarchical clustering, and the second object is the name of your plot file.

(iv) **[0.375 points]** Examine each of the three clustering graphs produced; we are trying to determine which clustering scheme produces the best clusters. We will define best to be the result that countries that are clustered together speak the same language, or languages. For example, Great Britain and Ireland should be clustered together because both countries contain a large percentage of people who speak EN (English).

Of the three graphs, which one produces what you would consider the best clusters? Please explain your answer as succinctly as you can. Put the answer to this question in the PDF you submit. Label the answer with “Q2.1(b)(iv) ...”

2.2 K-Means clustering

You are given a dataset in the file s1.csv on Canvas. Perform k-means clustering on that dataset by answering the questions below. Before answering the questions, plot the dataset and observe it visually. You DO NOT have to submit this plot.

(a) **[0.1 point[** Before using for modeling, should you standardize this dataset? Your function should return a string that contains one of: “yes”, “no”.

(b) **[0.2 points]** How many clusters does the silhouette method tell you are present in the data? Your function should return a scalar. (DO NOT submit the silhouette plot.)

(c) **[0.2 points]** How many clusters does the wss method tell you are present in the data? Your function should return a scalar. (DO NOT submit the wss plot.)

(d) **[0.5 points]** After consulting the output of 2.2(b) and 2.2(c) and visually examining the plot of the dataset, make a determination on how many clusters are depicted in the plot of the dataset. Run k-means using the number of clusters you determined. Plot the result from k-means to examine them visually, but DO NOT turn in the plot for submission. Your function should return the k-means object.

(e) **[1 point]** Looking at the plot in 2.2(d), comment on the clusters created by k-means. Indicate whether the clustering is satisfactory or not by justifying your answer in 1-3 sentences. Put the answer to this question in the PDF you submit. Label the answer with “Q2.2(e) ...”

2.3 DBSCAN clustering

Use the same s1.csv dataset that you used in question 2.2 for DBSCAN.

(a) **[0.1 point[** Before using for modeling, should you standardize this dataset? Your function should return a string that contains one of: “yes”, “no”.

(b) **[0.5 points]** Based on the problem description, what do you think will be an appropriate value for the size of the neighborhood set (i.e., MinPts parameter)? Your function should return a scalar representing the size of the neighborhood set.

(c) **[2 points]** In order to find the value of ε (eps), we need to calculate the average distance of every point to its *k* nearest neighbors. Set the value of *k* to be the result you obtained in 2.3(b). Then, using this value determine what the correct value for ε should be by creating a scree/elbow plot as shown in the lecture. Your function should return a scalar representing ε.

Note: To get the optimal value of you will have to try various values suggested by the scree/elbow plot to see how the cluster assignment changes. You may want to do a grid search across representative values of ε. To create the plots, use the following code:

```
df <- read.csv(…)
db <- fpc::dbscan(…)
plot(df, col=db$cluster+1, …)
```

Note: You DO NOT have to turn in the plots. They are for your reference only.

(d) **[0.4 points]** Comment, in 1-2 sentences, on the results of clustering after you have chosen MinPts and an optimal ε. Are all clusters well-separated? Are the noise points well identified? Put the answer to this question in the PDF you submit. Label the answer with “Q2.3(d) ...”