CS 422: Data Mining

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

Fall 2021: Homework 8 (10 points)

Due date: Sunday, Nov 21 2021 11:59:59 PM Chicago Time

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 (2 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 7, 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 Please label your answers clearly, see Homework 0 R notebook for an example (Homework 0 R notebook is available in "Blackboard → Assignment and Projects → Homework 0".) Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points per problem below.

2.1 K-means 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 **K-means** clustering on file19.txt on the above web page. This file contains a multivariate mammals dataset; there are 9 columns and 66 rows.

(a) Data cleanup (1 point divided evenly by components below)

(i) Think of what attributes, if any, you may want to omit from the dataset when you do the clustering. Indicate all of the attributes you removed before doing the clustering.

(ii) Does the data need to be standardized? (Briefly, using 1-2 sentences, support your answer.)

(iii) You will have to clean the data to remove multiple spaces and make the comma character the delimiter. Please make sure you include your cleaned dataset in the archive file you upload.

(b) Clustering (3 points divided evenly by components below)

(i) Determine how many clusters are needed by running the WSS or Silhouette graph. Plot the graph using fviz_nbclust().

(ii) Once you have determined the number of clusters, run k-means clustering on the dataset to create that many clusters. Plot the clusters using fviz_cluster().

(iii) How many observations are in each cluster?

(iv) What is the total SSE of the clusters?

(v) What is the SSE of each cluster?

(vi) Perform an analysis of each cluster to determine how the mammals are grouped in each cluster, and whether that makes sense? Act as the domain expert here; clustering has produced what you asked it to. Examine the results based on your knowledge of the animal kingdom and see whether the results meet expectations. Provide me a summary of your observations.

Hint: to get the indices of all animals in cluster 1, you would execute:

```
> which(k$cluster == 1)
```

assuming *k* is the variable that holds the output of the kmeans() function call.

2.2 dbscan clustering

Read in the dataset s1.csv uploaded in Blackboard. (s1.csv is extracted from "s1.txt", **Clustering Basic Benchmark**, P. Fränti and S. Sieranoja, http://cs.joensuu.fi/sipu/datasets/). S1 is a set of Gaussian clusters. There are 5,000 observations of two dimensions in the dataset.

(a) **[0.50 pts]** Do you think it is necessary to standardize the dataset? Justify your answer.

(b) **[0.10 pts]** (i) Plot the dataset.
(ii) **[0.10 pts]** Describe in 1-2 sentences what you observe (visually) in the plot: how many clusters do you see? Are they well-separated?

(c) Let's see how many clusters K-Means finds.
(i) **[0.10 pts]** Using the "wss" method, draw the scree plot for the optimal number of clusters.
(ii) **[0.10 pts]**Using the "silhouette" method, draw the scree plot for the optimal number of clusters.
(iii) **[0.50 pts]** What do you think is the appropriate number of clusters if we were to use K-Means clustering on this dataset?

(d) **[0.05 pts]** (i) Using the answer to (c)(iii), perform K-Means clustering on the dataset and plot the results.
(ii) **[0.50 pts]** Comment on how K-Means has clustered the dataset. **(1-2 sentences.)**

(e) We will now perform dbscan on this dataset.
(i) **[0.05 pts]**What is the value of MinPts that you think is reasonable for this dataset? Why?
(ii) **[2.00 pts]** 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 (e)(i). Then, using this value determine what the correct value for ε should be. (Hint: Look at the online manual page for the function kNNdistplot()).

Using the scree plot from kNNdistplot(), you should find the best value of ε that clusters the dataset into the expected number of clusters determined in (c)(iii). To do this, perform a grid search on ε, and for each value of

ε, run dbscan algorithm and visualize the clustering results. (You can do this manually in the R REPL and find the best value for ε, you do not need to write a loop.)

Using the best value of ε, plot the results of the dbscan algorithm on the dataset and state how many clusters you see in the plot in the form below:

```
At minPts = ___, eps = ____, there are ____ clusters.
```