

CS 422: Data Mining

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

Spring 2022: Homework 8 (10 points)

Due date: Tuesday, Apr 19 2022 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 (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, 11.

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 (6 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.