

CS 422: Data Mining

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

Spring 2018: Homework 3 (10 points)

Due date: Sunday, April 08, 2018 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.

This is the final version of the homework.

1. Exercises (2 points divided evenly among the questions) Please submit a PDF file containing answers to these questions. Any other file format will lead to a loss of 0.5 point. Non-PDF files that cannot be opened by the TAs will lead to a loss of 2 points.

1.1 Tan, Chapter 8

Exercise 2, 6, 11, 12, 16.

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 Problem 1: K-means clustering (3 points divided evenly among the components)

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

(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?

(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

- (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? For example, 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 Problem 2: Hierarchical clustering (2 points divided evenly among the components)

For this problem, you will use the “Languages spoken in Europe” dataset available from the Hartigan datasets. This dataset is available at <https://people.sc.fsu.edu/~jburkardt/datasets/hartigan/file46.txt>. Read this file into R. Note that you will have to pre-process the file since it is not, strictly speaking, a CSV file. Also make sure that the first column is recognized as a row label. (Hint: See the help on `read.csv()` and look at the `row.names` parameter.) Recognizing the first column as a row label is important because we want the country names to be printed as labels in the dendograms.

- (a) Run hierarchical clustering on the dataset using `factoextra::eclust()` method. Run the clustering algorithm for three linkages: single, complete, and average. Plot the dendogram associated with each linkage using `fviz_dend()`. Make sure that the labels (country names) are visible at the leafs of the dendogram.
- (b) Examine each graph produced in (a) and understand the dendrogram. Notice which countries are clustered together as two-singleton clusters (i.e., two countries clustered together because they are very close to each other in the shared languages/s). For **each** linkage method, list all the two singleton clusters. For instance, {Great Britain, Ireland} form a two-singleton cluster since they share English as a common language.
- (c) Italy is clustered with a larger cluster in the single and average linkage, whereas in complete linkage it is clustered with a smaller cluster. Which linkage strategy do you think accurately reflects how Italy should be clustered? (Hint: Look at the raw data.) Justify your answer in 1-2 sentences.
- (d) Let's pick a hierarchical cluster that we will call *pure*, and let's define *purity* as the linkage strategy that produces the most two-singleton clusters. Of the linkage methods you examined in (b), which linkage method would be considered *pure* by our definition?
- (e) Using the graph corresponding to the linkage method you chose in (d), at a height of about 125, how many clusters would you have?
- (f) Now, using the number of clusters you picked in (e), re-run the hierarchical clustering using the three linkage modes again, except this time through, specify the number of clusters using the *k* parameter to

`factoextra::eclust()`. Plot the dendrogram associated with each linkage using `fviz_dend()`. Make sure that the labels (country names) are visible at the leafs of the dendrogram.

(g) For each cluster obtained by the value of k used in (f), print the Dunn and Silhouette width using the `fpc::cluster.stats()` method. Take a look at the help (or manual) page for `fpc::cluster.stats()` and see what is the name of the return list component that contains the Dunn index and the average Silhouette width.

(h) From the three clusters in (g), which is the best cluster obtained if you consider the Dunn index only?

(i) From the three clusters in (g), which is the best cluster obtained if you consider the Silhouette width only?

2.3 Problem 3: K-Means and PCA (3 points divided evenly among the components)

HTRU2 is a data set which describes a sample of pulsar candidates collected during an astronomical survey. More information on HTRU is provided on the UCI Machine Learning Repository (see <https://archive.ics.uci.edu/ml/datasets/HTRU2>). The dataset consists of 17,898 observations in 8 dimensions, with the 9 attribute being a binary class variable (0 or 1). The dataset is available to you on Blackboard; it is highly recommended that you read the UCI Machine Learning Repository link given above to get more information about the dataset.

(a) Perform PCA on the dataset and answer the following questions:

(i) How much cumulative variance is explained by the first two components?

(ii) Plot the first two principal components. Use a different color to represent the observations in the two classes.

(iii) Describe what you see with respect to the actual label of the HTRU2 dataset.

(b) We know that the HTRU2 dataset has two classes. We will now use K-means on the HTRU2 dataset.

(i) Perform K-means clustering on the dataset with $k = 2$. Plot the resulting clusters.

(ii) Provide observations on the shape of the clusters you got in (b)(i) to the plot of the first two principal components in (a)(ii). If the clusters are similar, why? If they are not, why?

(iii) What is the distribution of the observations in each cluster?

(iv) What is the distribution of the classes in the HTRU2 dataset?

(v) Based on the distribution of the classes in (b)(iii) and (b)(iv), which cluster do you think corresponds to the majority class and which cluster corresponds to the minority class?

(vi) Let's focus on the larger cluster. Get all of the observations that belong to this cluster. Then, state what is the distribution of the classes within this large cluster; i.e., how many observations in this large cluster belong to class 1 and how many belong to class 0?

(vii) Based on the analysis above, which class (1 or 0) do you think the larger cluster represents?

(viii) How much variance is explained by the clustering?

(ix) What is the average Silhouette width of both the clusters?

(x) What is the per cluster Silhouette width? Based on this, which cluster is *good*?

(c) Perform K-means on the result of the PCA you ran in (a). More specifically, perform K-means on the first two principal component score vectors (i.e., `pca$x[, 1:2]`). Use $k = 2$.

(i) Plot the clusters and comment on their shape with respect to the plots of a(ii) and b(i).

(ii) What is the average Silhouette width of both the clusters?

(iii) What is the per cluster Silhouette width? Based on this, which cluster is *good*?

(iv) How do the values of c(ii) and c(iii) compare with those of b(ix) and b(x), respectively?