

## CS 422: Data Mining

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

### Spring 2019: Homework 3 (10 points)

**Due date: Monday, April 15 2019 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 assignment is complete. (Last change: Apr-10)**

**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 7 (Cluster Analysis: Basic Concepts and Algorithms)**

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 among the components as shown below)**

Download the BuddyMove dataset from UCI Machine Learning repository; the dataset is available at <https://archive.ics.uci.edu/ml/datasets/BuddyMove+Data+Set>. The dataset consists of 249 observations in six dimensions; each observation corresponds to a user (a reviewer) who has submitted a certain number of reviews corresponding to his or her interests.

**Your overall goal is to use k-means clustering to determine if you can cluster like users together.** We will assume that the definition of “like users” is a set of users that have the same interests. For example, User 28 clearly enjoys visiting religious places and shopping. So one possible cluster could correspond to users that enjoy the same things. Beyond reaching this goal, I have no explicit steps for you to follow; I want you to be as creative as possible in achieving this goal. I want you to think whether transforming the data will help in any way or whether you can use the data in the form it is in. Working with User 28, the user submitted 53 reviews for Nature; is 53 reviews a lot? Average? Above average? Think :-).

(a) You may choose to use `fviz_nbclust()` to determine the optimum number of clusters; if so, feel free to use it and plot the graph using `fviz_nbclust()`.

- (b) 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()`.
- (c) How many observations are in each cluster?
- (d) What is the total SSE of the clusters?
- (e) What is the SSE of each cluster?
- (f) Perform an analysis of each cluster to determine how the users 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 gained from playing around with the dataset, see whether the results meet expectations. Provide me a summary of your observations.

Hint: to get the indices of all users 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)

The aim of this problem is to observe how hierarchical clustering works to cluster like users together. We will use the same dataset as Problem 2.1. **Scale the data before use.**

**This is important:** make sure that the first column is recognized as a row label when you read the dataset in. (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 user IDs to be printed as labels in the dendograms.

For this problem, you will use a sampled subset of the above dataset consisting of 50 random observations. To get this subset, **set the seed to 1122** and use the `dplyr::sample_n()` function (see the R help page on how to use it). Failure to set the seed to 1122 will result in a different sample, and in such a case you will not get full credit for the assignment.

- (a) Run hierarchical clustering on the dataset using `factoextra::eclust()` method, use `k=1` parameter to force a single cluster. Run the clustering algorithm for three linkages: single, complete, and average. Plot the dendrogram associated with each linkage using `fviz_dend()`. Make sure that the labels (User IDs) are visible at the leafs of the dendogram.
- (b) Examine each graph produced in (a) and understand the dendrogram. Notice which users are clustered together as two-singleton clusters (i.e., two users are clustered together because they are very close to each other in the attributes they share). For **each** linkage method, list all the two-singleton clusters. For instance, {User 43, User 35} form a two-singleton cluster in the average linkage method since they share a lot of the same characteristics.
- (c) We will now determine how many clusters to form. Let's pick a hierarchical cluster that we will call *pure*, and let's define *purity* as the linkage strategy that produces the **least** two-singleton clusters. Of the linkage methods you examined in (b), which linkage method would be considered *pure* by our definition?
- (d) Using the graph corresponding to the linkage method you chose in (c), draw a horizontal line at a height of 1.7. How many clusters would you have?

(e) For the number of clusters you determined in (d), re-run hierarchical clustering across the three linkage strategies (single, average, and complete) with the value of  $k$  being the number of clusters you determined in (d). For each linkage strategy, find out its Silhouette index.

Download **NbClust** package. This package proposes a best clustering scheme using different methodologies for determining clusters.

(f) For each linkage strategy, determine the number of clusters that **NbClust()** suggests. Take a look at the **method=** parameter in **NbClust()** and pass in the linkage method, e.g.: **NbClust(data, method="single")**.

(g) For the number of clusters you determined for each linkage in (f), find out its Silhouette index.

(h) You used two strategies to cluster: one was in part (c) where you used purity to define clusters, and the other strategy was using **NbClust()**. Between these two strategies, pick the linkage that is best, as defined by the Silhouette index. Comment on which of the two strategies comes closest to your expectations.

### 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 smaller version of the dataset (10,000 observations) is available to you on Blackboard. **Be aware that even running this small version of the dataset in some of the questions below will take time if your laptop/machine does not have enough RAM. A Windows 7 dual-core system with 4GB RAM has been known to take 9 hours to process the small dataset. So, please start early and make sure you leave enough time for the program to run if your machine does not have enough resources.**

It is also highly recommended that you read the UCI Machine Learning Repository link given above to get more information about the dataset.

**Use the smaller version of the HTRU2 dataset to answer all the questions below.**

**(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 **centers = 2**, and **nstart = 25** (otherwise your answers will not match and you will not get points). 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?