BUAN 6357 (Johnston)
Homework 2 (20230119)
Code Due: 4 February 2023 (6:00PM)
Part B Due: 5 February 2023 (11:59PM)

Points available: 60

This assignment compares the use of kmeans() and hclust(). Data for training and testing sets are provided as well as a seed for the RNG.

You will only need the “tidyverse” and “data.table” packages for this assignment. Do not use any other “require()” or “library()” statement in your code. Any code which invokes install.packages() will be given a score of 0. Any instance of more than 1 code submission in the queue will result in only the last instance being run.

The first commands of your code MUST be:

setwd(“c:/data/BUAN6357/HW_2”); source(“prep.txt”, echo=T)

and the last command of your code MUST be:

source(“validate.txt”, echo=T)

Be careful with the quote characters as they must ALL be the same at the beginning and end of a string.(Use the single or double quote character from the key next to “Enter”.) Inclusion of these lines is required BEFORE your code will be tested.

1. The training and testing data can be loaded into you workspace via the fread() function. The 2 files to be loaded are available in UTDbox>HW_assignments as HW_2_train.csv and HW_2_test.csv. Use the training set to determine the appropriate number of clusters using kmean() and again using hclust().
2. The RNG seed is 893571057

3. The hclust() should use a squared Euclidean distance as the function input and will require that you calculate the penalty function values separately.
4. The penalty function to be used for both kmeans() and hclust() in this assignment is the Total Within-Group Error Sum of Squares.

Submit the code to eLearning as an ASCII file, with the ".txt" extension, which can be copied directly into R.

You may submit this assignment as many times as needed until you get full credit. Only the last score for each portion of a HW (A or B) will be used in calculation of the course grade.

Deliverables (all names case as shown) :

1. seed (type: numeric vector) RNG seed
2. train (type: data.table) training set data
3. test (type: data.table) testing set data
4. kmTWSS (type: numeric vector) cluster scenario specific TWSS for kmeans() from training set
5. hcTWSS (type: numeric vector) cluster scenario specific TWSS for hclust() from training set
6. hcObj (class: hclust) merge tree for training set

Part B of HW 2 will ask questions related to the interpretation of your results and may require additional programming.

Notes:

- Use default parameter settings on hclust(), specifically, use ***method="complete"*** which is the default option. We will look at the other clustering options later in the course.
- Use default parameter settings on dist(), specifically, use ***method="euclidean"*** which is the default option.
- hclust() provides only a merge list in the returned object. You need to use cutree() to retrieve a cluster membership vector for each clustering scenario (that is, where you want to investigate the results of using a specific number of clusters, as was done when calling kmeans() multiple times).

- hclust() does ***not*** provide you with a penalty function value for each scenario. You will need to use the cluster membership vector to calculate this on your own. While this may appear inconvenient at this time, it provides much more flexibility in the clustering process, as does the use of the dist() function, which has its own set of options. Notice that we are already expanding the options available through dist() by using a transformation of it values.
- kmeans() have a parameter for the number of starting attempts to find a good initial allocation of observations to clusters. Use "nstart=5" for this assignment.
- Run the number of clusters scenarios from 1 to 10 for Kmeans and hclust().