Stat 437 HW3

Your Name (Your student ID)

General rule

You must complete both Conceptual and Applied exercises. Please show your work and submit your computer codes in order to get points. Providing correct answers without supporting details does not receive full credits. Please upload your answers to the course space. This HW covers

- K-means clustering
- Hierarchical clustering

For an assignment or project, you DO NOT have to submit your answers or reports using typesetting software. However, your answers must be well organized and well legible for grading. Please upload your answers in a document to the course space. Specifically, if you are not able to knit a .Rmd/.rmd file into an output file such as a .pdf, .doc, .docx or .html file that contains your codes, outputs from your codes, your interpretations on the outputs, and your answers in text (possibly with math expressions), please organize your codes, their outputs and your answers in a document in the format given below:

```
Problem or task or question ...
Codes ...
Outputs ...
Your interpretations ...
```

It is absolutely not OK to just submit your codes only. This will result in a considerable loss of points on your assignments or projects.

Conceptual exercises

1. Consider the K-means clustering methodology.

1.1) Give a few examples of dissimilarity measures that can be used to measure how dissimilar two observations are. What is the main disadvantage of the squared Euclidean distance as a dissimilarity measure?

1.2) Is it true that standardization of data should be done when features are measured on very different scales? Is it true that employing more features gives more accurate clustering results? Is it true that employing standardized observations gives more accurate clustering results than employing non-standardized ones? Explain each of your answers.

1.3) Take $K = 2$. Provide the loss function that K-means clustering tries to minimize. You need to provide the definition and meaning of each term that appears in the loss function.

1.4) What is the "centroid" for a cluster? Is the algorithm, Algorithm 10.1 on page 388 of the Text (which is also provided in the lecture slides), guaranteed to converge to the global minimum of the loss function? Why or why not? What does the argument `nstart` refer to in the command `kmeans`?

Why is `nstart` suggested to take a relatively large value? Why do you need to set a random seed by `set.seed()` before you apply `kmeans`?

1.5) Suppose there are 2 underlying clusters but you set the number of clusters to be different than 2 and apply `kmeans`, will you have good clustering results? Why or why not?

1.6) Is the true number K_0 of clusters in data known? When using the command `clusGap` to estimate K_0, what does its argument `B` refer to?

2. Consider hierarchical clustering.

2.1) What are some advantages of hierarchical clustering over K-means clustering? What is the relationship between the dissimilarity between two clusters and the height of these clusters in the dendrogram that represents a bottom-up tree?

2.2) Explain what it means by saying that "the clusters obtained at different heights from a dendrogram are nested". If a data set has two underlying clustering structures that can be obtained by two different criteria, will these two sets of clusters necessarily be nested? Explain your answer.

2.3) Why is the distance based on Pearson's sample correlation not effected by the magnitude of observations in terms of Euclidean distance? What is the definition of average linkage? Why are average linkage and complete linkage preferred than single linkage in practice?

2.4) What does the command `scale` do? Does `scale` apply row-wise or column-wise? When `scale` is applied to a variable, what will happen to the observations of the variable?

2.5) What is `hclust$height`? How do you find the height at which to cut a dendrogram in order to obtain 5 clusters?

2.6) When creating a dendrogram, what are some advantages of the command `ggdendrogram{ggdendro}` over the R base command `plot`?

Visualizing clustering results as, e.g., done by Example 1 in "LectureNotes3_notes.pdf".

Applied exercises

3. Please refer to the NYC flight data `nycflights13` that has been discussed in the lecture notes and whose manual can be found at https://cran.r-project.org/web/packages/nycflights13/index.html. We will use `flights`, a tibble from `nycflights13`.

Select from `flights` observations that are for 3 `carrier` "UA", "AA" or "DL", for `month` 7 and 2, and for 4 features `dep_delay`, `arr_delay`, `distance` and `air_time`. Let us try to see if we can use the 4 features to identify if an observation belongs a specific carrier or a specific month. The following tasks and questions are based on the extracted observations. Note that you need to remove `na`'s from the extracted observations.

3.1) Apply K-means with $K = 2$ and 3 respectively but all with `set.seed(1)` and `nstart=20`. For $K = 3$, provide visualization of the clustering results based on true clusters given by `carrier`, whereas for $K = 2$, provide visualization of the clustering results based on true clusters given by `month`. Summarize your findings based on the clustering results. You can use the same visualization scheme that is provided by Example 2 in "LectureNotes3_notes.pdf". Try visualization based on different sets of 2 features if your visualization has overlayed points.

3.2) Use `set.seed(123)` to randomly extract 50 observations, and to these 50 observations, apply hierarchical clustering with average linkage. (i) Cut the dendrogram to obtain 3 clusters with leafs annotated by `carrier` names and resulting clusters colored distinctly, and report the corresponding height of cut. (ii) In addition, cut the dendrogram to obtain 2 clusters with leafs annotated by `month` numbers and resulting clusters colored distinctly, and report the corresponding height of cut. Here are some hints: say, you save the randomly extracted 50 observations into an object `ds3sd`, for these observations save their `carrier` names by keeping their object type but save `month` numbers as a `character` vector, make sure that `ds3sd` is a `matrix`, transpose `ds3sd` into `tmp`, assign to `tmp` column names with their corresponding carrier names or month numbers, and then transpose `tmp` and save it as `ds3sd`; this way, you are done assigning cluster labels to each observation in `ds3sd`; then you are ready to use the commands in the file `Plotggdendro.r` to create the desired dendrograms.