Student Name: Brandon Russell

Submission Date: 7/15/2018

DBST 667 – Data Mining

Dr. Irene Tsapara

**Week 8 Individual Exercise**

**Deliverables:** Two Files: (1) Submit this lab report with answers to all questions including output screenshots into the ‘Individual Exercises Week 8’ assignment folder. (2) Submit an R script that contains all commands with comments that briefly describe each commands purpose.

**Grading: This exercise is worth 2% of the course grade.** All questions must be answered in your own words with any paraphrased references properly cited using in-text citations and a reference list as needed. In addition, grammatical and spelling errors may affect the grade.

**Part 2** – **Run an exercise on the *Vehicle Solhouettes* dataset from vehicle.csv, completing this report and providing the commands, output screenshots, and discussion/interpretation as requested. Ensure that all commands are saved in this report AND in an R script.**

**For Reference:** [**UCI Machine Learning Repository: Vehicle Silhouettes**](http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29)

1. **Introduction:** 
   1. **Based on what you have learned this week about k-means clustering, provide a one-paragraph masters-level response describing what you anticipate that the kmeans method will accomplish for the Vehicle Silhouettes data? Be specific about the behavior and output structure of k-means models.**
2. **Data Pre-Processing: Load the Vehicle Silhouettes data into R Studio using the read.csv command (*do not use File > Import Dataset > From CSV in the R Studio GUI as this uses read\_csv() resulting in significant different variable types!!!*).**
   1. **Make a copy of the loaded Vehicle Silhouettes data you just imported and name the copy ‘myvehicle’. Keep the original import as you will need both the original and copy to complete this report. Include the command demonstrating this step below.**

**Command: > myvehicle <- vehicle**

* 1. **Remove the variable class from ‘myvehicle’. Include the command and answer to the question below.**

**Command: > myvehicle$Class <- NULL**

**Why do we need to remove the class variable as part of the data preprocessing steps for k-means clustering?**

**The kmeans functions in R requires the use of numeric data. “Class” was a factor. It was also a label, which is simple not needed with clustering.**

* 1. **Run the scale() function on ‘myvehicle’. Include the command and answer to the question below. (*Note: This command is NOT part of your tutorial. Consult the function help and use the default arguments. Hint: scale() is a function that outputs its results. You MUST save the scaled output back to the original ‘myvehicle’.***

**Command: > myvehicle <- scale(myvehicle)**

**Why must we scale data as part of the data preprocessing steps for k-means clustering?**

* 1. **What additional data preprocessing steps (if any) did you need to execute? Include the command(s) and output screenshot below.**

**Command(s): >**

**Output:**

1. **K-Means Clustering – Running the Method (*Hint: Record your results with k=4 in the table in part f)*:**
   1. **Run ‘set.seed(12345)’ and then run the kmeans method with k=4 and store the output to a variable named ‘kc’. Include the command, output screenshot, and discuss the input parameters you used.**

**Command: >**

**Output:**

**Discussion:**

* 1. **Enter ‘kc’ at the prompt. Provide the output below and then answer the following questions:**

**Output:**

**How many instances are in each cluster?**

**What information does the cluster means section provide and how were those numbers obtained?**

**What is the clustering vector?**

**What is the sum of squares by clusters and what does it mean?**

* 1. **Run the ‘kc$iter’ command. Include the command, output screenshot, and explain what the output shows.**

**Command: >**

**Output:**

**Discussion:**

1. **K-Means Clustering – Evaluate the Model:** 
   1. **Build the cross-tabulation to compare how the method clustered the vehicles from ‘myvehicle’ to the actual vehicle class from your original import. Include the command, output screenshot, and answer the following questions:**

**Command: >**

**Output:**

**What is the dominant vehicle class in each cluster?**

**What is the dominant cluster for each vehicle class?**

**What percentage of vehicles were clustered in agreement with the actual class?**

1. **K-Means Clustering – Cluster Visualization:**
   1. **Run the ‘clusplot(kc)’ function to visualize your model. Modify the plot appearance to make your visualization clear and easy to interpret. Unlike previous exercises, your visualization will now be evaluated on clarity and aesthetics in addition to the standard command, output, and interpretation evaluation. Include the full command, output screenshot (zoomed in), and a one-paragraph, masters-level response with your interpretation of your plot.**

**(*Hint: Your interpretation should discuss all of the visualized clusters and should begin to address specific observations (data points) within each that warrant discussion.)***

**Command: >**

**Output:**

**Interpretation:**

1. **K-Means Clustering – Experiment with Different K Values (3 Runs Summarized):**
   1. **Completely fill in the table below documenting the results of your experimentation with modifying the k value. You may use any k value other than 4 that is greater than 0. You do not need to provide any commands or output screenshots in this report. However, you will be evaluated on these commands being present in your R script!**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **k=** | **Number of Instances in Each Cluster** | **Between Clusters Sum of Squares** | **Within Clusters Sum of Squares** | **Number of Iterations** |
| **4** |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

* 1. **What effect do you observe that modifying the k values has on the method results? Provide a one-paragraph, masters-level response below:**
  2. **What is an ideal value of k for the Vehicle Silhouettes data? This is a subjective and open-ended question. Challenge yourself and come up with a creative and well-supported answer for which value you believe is ideal. Provide a one-paragraph, masters-level response below:**

1. **Summary:**
   1. **What differences between k-means clustering and classification methods did you observe? Provide a one-paragraph, masters-level response.**
   2. **(Not graded) Which part of this exercise did you find the most challenging and what steps did you take to resolve the challenge?**

References