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.**

**The k-means algorithm will assess the center value for each of our attributes. After doing so it will randomly select a cluster center for each k and begin assigning observations based on their attributes Euclidean distance to the cluster center values. After running through a single iteration, it will assess the variation within clusters and then attempts to improve the variance by running another iteration with different cluster mean values. This is repeated until no more improvements can be made (Han, Kamber, & Pei, 2011). The output will provide us a lot of information, such as the total number of instances assigned to a cluster, a table of every cluster an observation was definitively assigned to, and other measurements to determine the accuracy of our clustering, i.e. between and within clusters sum of squares (Williams, 2014).**

1. **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 function in R requires the use of numeric data. “Class” was a factor. It was also a label, which is simply not needed for 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)**

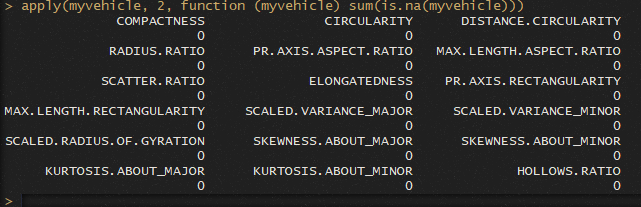
**Why must we scale data as part of the data preprocessing steps for k-means clustering?**

**Scaling our data is important for balancing the variance in the attributes used for clustering. For example, the “Scaled.Variance\_Minor ranges from 184 to 1018, while the “Scaled.Radius.Of.Gyration” only has a range of 109 to 268. With this imbalance the weight of one attribute might suppress the others. When these two are scaled their values are between -1.45 to 3.27 and -2 to 2.9, respectively, which brings them within a smaller variance without diminishing the impact of the original datapoints.**

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

**Command(s): > apply(myvehicle, 2, function (myvehicle) sum(is.na(myvehicle)))**

**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: > kc<-kmeans(myvehicle, 4)**

**Output:**

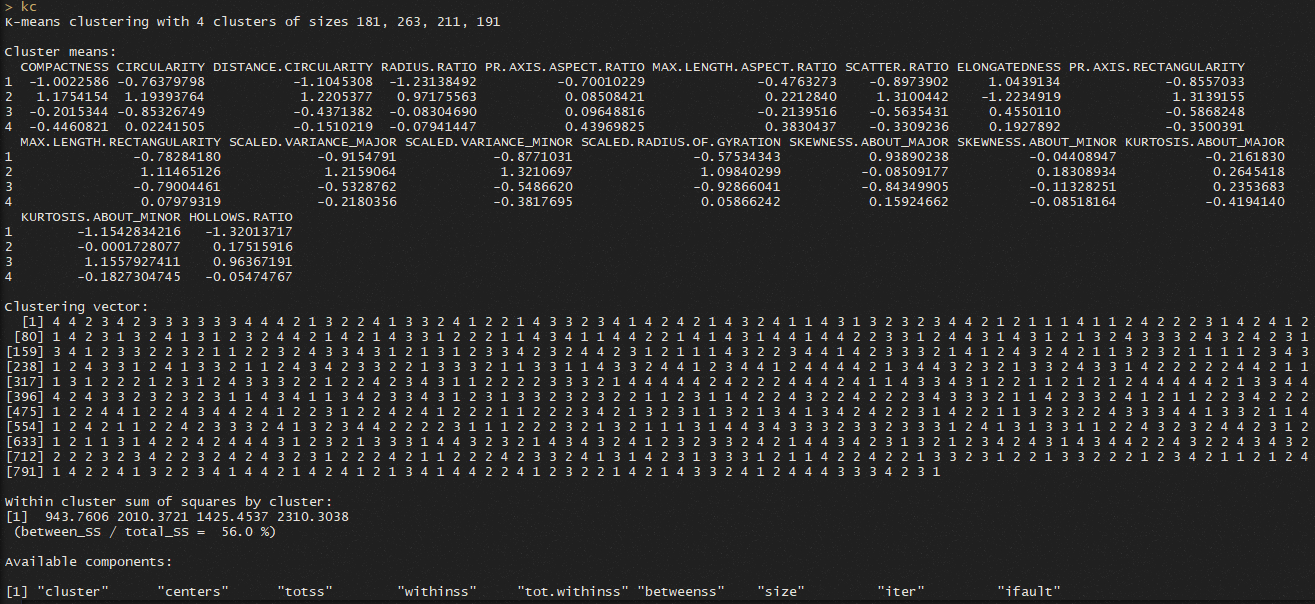


**Discussion:**

**The set.seed command simply ensures we are randomizing our data. The kmeans command is implementing our clustering. The first input is the dataset. The second input is the number of clusters we want the algorithm to build.**

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

**Output:**



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

**181, 263, 211, 191**

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

**This section lists every attribute for an instance, and then for each attribute it provides the mean value of that attribute within each cluster for that specific instance. After the kmeans algorithm runs through its procedure of determining means and calculating distances until the optimal mean is discovered, in the last iteration these values are collected.**

**What is the clustering vector?**

**This section lists every instance in the dataset, the value associated with the instance is the cluster it was assigned to. For example, the first instance belongs to cluster 4.**

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

**The sum of squares by cluster is telling us the sum of the squared distances between instances assigned to a cluster to the mean of the cluster. In other words, this is a measure of the overall variance of the cluster (Williams, 2014).**

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

**Command: > kc$iter**

**Output:**

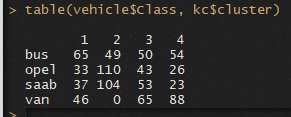


**Discussion: The “iter” function of kmeans tells us the number of iterations the algorithm had.**

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: > table(vehicle$Class, kc$cluster)**

**Output:**



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

**Cluster 1 – “Bus” > 65**

**Cluster 2 – “Opel” > 110**

**Cluster 3 – “Van” > 65**

**Cluster 4 – “Van” > 88**

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

**“Bus” – Cluster 1 > 65**

**“Opel” – Cluster 2 > 110**

**“Saab” – Cluster 2 > 104**

**“Van” – Cluster 4 > 88**

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

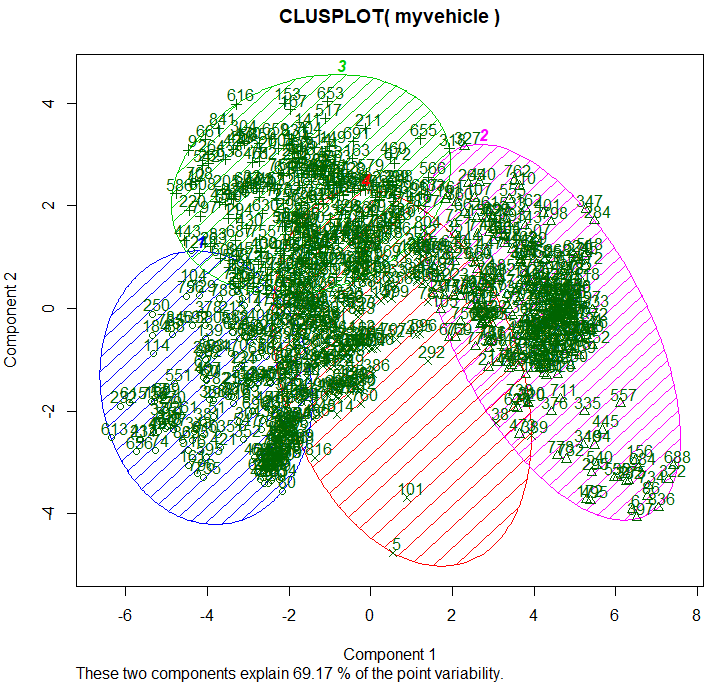
**(65+110+104+88)/846 = 0.434 = 43%**

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: > clusplot(myvehicle, kc$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)**

**Output:**



**Interpretation:**

**In the above figure, we see a plot of our vehicle dataset within the defined clusters. Each cluster is labeled and represented by a color: cluster 1 is blue, 2 is purple, 3 is green, and 4 is red. All the numbers within each cluster represent all the instances of the dataset and which cluster they belong to. For example, we can clearly see instance 114 belongs to cluster 1, instance 653 belongs to cluster 3, instance 101 belongs to cluster 4, and instance 557 belongs to cluster 2. The overlapped areas highlight an inability to definitively classify an instance, an indication of the amount of error. This function makes use of principal component analysis to determine the components to plot the data with for optimal results, and here we can see we are able to visualize 69.17% of our data cluster variability with the chosen characteristics (Han, Kamber, & Pei, 2011).**

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** | **181**  **263**  **211**  **191** | **8520.11** | **6663.868** | **4** |
| **3** | **277**  **540**  **29** | **6785.479** | **7299.396** | **2** |
| **2** | **551**  **295** | **6251.343** | **8958.657** | **1** |
| **1** | **846** | **-1.818989e-12** | **15210** | **1** |

* 1. **What effect do you observe that modifying the k values has on the method results? Provide a one-paragraph, masters-level response below:**

**There are several interesting reflections we can note from the table above as we reduced our k value from 4 to 1. First, the between clusters sum of squares decreased. This makes sense as it is measuring the distance between clusters, therefore the fewer clusters the smaller this number should be. Secondly, we can note the within clusters sum of squares increases as we decrease our k value. This is a measure of distance between observations and their cluster center, therefore the bigger the cluster is the more chance there is for observations to be further from the center. Finally, we can see the number of iterations had an exponential decrease. Based on this information we can conclude that decreasing the value of k is technically more efficient and quicker in determining cluster information, though there is a tradeoff with accuracy. Optimally we would want our between clusters sum of squares to be large, while the within sum of squares would reduce in size to provide the best accuracy in our clustering. It would become a decision when using these numbers between the level of granularity we want and the amount of complexity we are willing to add to our function in determining an acceptable k value (Williams, 2014).**

* 1. **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:**

**In the case of testing values 1 through 4, I would say having a k value of 3 is most optimal. In this value, we had a between clusters sum of squares of 6,785 and a within cluster sum of squares value of 7,299. While having a k value of 4 provided better numbers, a k value of 3 only required 2 versus 4 iterations. I would argue splitting our computation in half for a slight decrease in cluster accuracy is acceptable for our needs. Additionally, moving down to a k value of 2 sees a significant increase in the within clusters sum of squares.**

1. **Summary:**
   1. **What differences between k-means clustering and classification methods did you observe? Provide a one-paragraph, masters-level response.**

**The most obvious difference between k-means clustering and classification methods, such as decision trees, is the class labels, or lack thereof. With k-means clustering, we use a form of distance measurement, such as Euclidean distance, from a created cluster center to determine if an observation belongs to a certain cluster, sometimes with the use of Principal Component Analysis (PCA). Of course, once an observation is assigned to a cluster we still cannot positively identify the observation, we can only learn what other observations it is like and why. This is considered an unsupervised learning algorithm. Classification, on the other hand, gives an actual class label to an observation. This is done by teaching the algorithm what defines a class label using a training dataset. With decision trees, for example, our algorithms use information gain or the Gini index to determine which attributes to use in a splitting decision to ultimately classify an observation. For example, with our vehicle dataset and clustering we might find vehicles with a similar distance to the mean value of “Max.Length.Aspect.Ratio” and “Compactness” attributes within the same cluster. However, with decision trees we might find the “Circularity” attribute has the highest weight in determining a bus or van label (Han, Kamber, & Pei, 2011).**

* 1. **(Not graded) Which part of this exercise did you find the most challenging and what steps did you take to resolve the challenge?**

**I initially did not fully understand the “iter” value, however after researching a few different online resources, I was able to figure it out.**

# References

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques.* Elsevier.

Williams, G. (2014, June 22). *Data Science with R Cluster Analysis*. Retrieved from OnePageR: https://onepager.togaware.com/ClustersO.pdf