NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp

Cluster Analysis

Question #1: K-Means

Read in the `[08] Protein.txt` dataset into your workspace. This dataset contains protein consumption information from 1973 on nine different food groups across 25 different European countries.

1. Use the following commands to read the data into your workspace appropriately and scale the variables:

```
protein = read.table("08 Protein.txt", sep = "\t", header = TRUE)
protein.scaled = as.data.frame(scale(protein[, -1]))
rownames(protein.scaled) = protein$Country
```

2. Create and interpret a scree-plot for the within-cluster variance for various values of K used in the K-means algorithm.
    a. Why might this graph indicate that K-means is not truly appropriate to model the data?
3. Create and store 5 different K-means solutions that run the algorithm only 1 time each. (**NB:** Use `set.seed(0)` so your results will be reproducible.)
4. Create and store 1 K-means solution that was selected from running the algorithm 100 separate times. (**NB:** Use `set.seed(0)` so your results will be reproducible.)
5. Plot the 6 different solutions from part 3 and 4 with:
    a. `Cereals` on the x-axis.
    b. `RedMeat` on the y-axis.
    c. Colors for the different cluster assignments.
    d. Labels for the total within-cluster variances.
6. Plot the solution from part 4 with:
    a. `Cereals` on the x-axis.
    b. `RedMeat` on the y-axis.

c. A label for the total within-cluster variance.
d. Points for the centroids of each cluster.
e. A horizontal line at 0.
f. A vertical line at 0.
g. Text listing the country for each observation in your dataset (instead of points), colored by the different cluster assignments. **Hint:** Use `type = "n"` when creating the `plot()`. Then, use the `text()` function in tandem with the `rownames()` function.

7. Interpret the clustering solution based on the graph you created in part 6.

Question #2: Hierarchical Clustering

Continue using the `[08] Protein.txt` dataset you already loaded into your workspace.

1. Calculate and store pairwise distances for each observation in the dataset.
2. Fit hierarchical clustering solutions using single, complete, and average linkage.
3. Visualize the dendrograms created in part 2.
    a. Give an argument as to why single linkage might not be good to use.
    b. Give an argument as to why complete linkage might be good to use.
4. Cut your complete linkage tree into 2 groups.
    a. Visualize the solution overlaid on top of the dendrogram.
    b. Interpret the clusters by aggregating across the median.
5. Cut your complete linkage tree into 5 groups.
    a. Visualize the solution overlaid on top of the dendrogram.
    b. Interpret the clusters by aggregating across the median.