Assignment #5: Clustering (4 pts)

Group Submission

Due: November 28 2:00 pm.

In this problem, we perform clustering on the states using the USArrests.csv data set. There are four dimensions:

- Murder: Murder arrests (per 100,000)
- Assault: Assault arrests (per 100,000)
- UrbanPop: Percent urban population
- Rape: Rape arrests (per 100,000)

i) Using hierarchical clustering with average linkage and Euclidean distance, cluster the states. Cut the dendrogram to obtain four clusters. List the states in each cluster.

ii) Perform K-means clustering with K=4 on the data set. Use the function table() to compare the results obtained in i) and ii). Do they return the same clustering result? For each K-means cluster, what is its within-cluster sum of square? How distinct each cluster is from other clusters?

iii) In K-means clustering, could you find the optimal value of K? Justify your answers. Perform K-means clustering using your selected K. List the states in each cluster.

Deliverables

1. Group submission. Each group submits one set of report and code. Please include a cover page on the report listing all team members' names.
2. Two files: R code and the report are submitted as two separate files to Blackboard. Screenshot of R code is not accepted.
3. The report should contain the answer to each question. No R raw outputs or software screenshot should be included in the report except plots.