CITS5508 Machine Learning
Semester 1, 2024
Lab sheet 9
(Not assessed)

You will develop Python code to perform clustering using principal components analysis (PCA). Certify that the presentation of your Python notebook is good and that you used the Markdown cells well. Make sure you properly format your plots and results. For instance, all your diagrams/plots should have proper axis labels and titles to help the reader understand what you are plotting. Use the lab sheets to learn how to improve the presentation of your notebook, as you will need this in the assessments.

A clustering analysis on the USArrests data

The `USArrests` data contains the statistics, in arrests per 100,000 residents, for three crime-related features (assault, murder and rape) for all 50 US states in 1973. An additional feature, UrbanPop, is also included and describes the percentage of the population living in urban areas.

The provided `USArrests.csv` data set contains 50 observations on 4 variables:

- Murder: murder arrests (per 100,000)
- Assault: assault arrests (per 100,000)
- Rape: rape arrests (per 100,000)
- UrbanPop: percent of the population living in urban areas

Tasks

1. Using the raw data, perform a hierarchical clustering with complete linkage and Euclidean distance to cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which cluster? Describe their characteristics.
2. Repeat Task 1 after scaling the variables to have zero mean and unit standard deviation. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled? Justify for your answer.
3. Perform PCA on the data. Now perform hierarchical clustering with complete linkage and Euclidean distance on the first two principal component score vectors rather than the raw data. Cut the dendrogram at a height that results in three distinct clusters. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (three colours for three clusters). Compare the group characteristics to the group characteristics obtained in Task 2.
4. Repeat the analysis of Task 3 using the K-means clustering (with K=3). That is, use the first two principal components score vectors as features and set the initial centroids of the K-means as the group means obtained from the hierarchical clustering on Task 3. Compare the results from the K-means clustering to the results from the hierarchical clustering of Task 3. Which one do you think provides a better result? For the K-means clustering, you should use the Euclidean distance and set `random_state` to "5508".