MTH6101 Introduction to Machine Learning

Laboratory week six

The intention of this laboratory is to do a second take on clustering with functions `kmeans` and `pam`, this second from the library `cluster`.

1. Open `RStudio` and create a new `R` script. Load library `cluster`.

2. Consider the data set `USArrests` of arrests statistics in states of the USA. Load it into a variable termed `X1`, taking care of centering and scaling the data. Perform `kmeans` clustering of `X1` for values of $k = 2, \ldots, 10$ and record the total sum of squares within clusters. Plot this quantity.

3. Plot the data using text labels and color using the clusters you have found (`$cluster`). To this end do a scatterplot with `type="n"` and then add labels using `text`. Clearly we need to select some variables for the plot so use `Assault` and `UrbanPop`. Interpret your result.

4. You will now redo all the analysis for the Principal Component scores of this data. To this end, redo PCA for the data `USArrests` (centered, scaled), select a number of components and store the PC Scores in variable `X2`. Then do cluster analysis using `kmeans` **on the scores**. Plot ESS for a selection of k between 2 and 10.

5. Do two scatter plots with text labels, one with the variables `Assault` and `UrbanPop` and the colors given by the clustering on PC you just did. The second plot is of `X2` which are PC scores, and the same coloring. Interpret the results.

6. Build `X3` to be the centered and unscaled `iris` data set (no fifth column). Using the medoid clustering method `pam` cluster with $k = 2, 3, 4, 5$ medoids and plot the clusters for sepal variables, coloring according to the clusters made (`$clustering`). Can you suggest some clusters?

7. Repeat the previous step for `X3` not just centered but scaled `iris` data set without the fifth column. Are clusters neater?

8. Repeat the previous `pam` clustering with `X4` which are first two PC scores of the centered and scaled `iris` data set (no fifth column). Can you determine how many clusters would be suitable?