Exercise Lesson02: KMeans

Machine Learning using R

Exercise 1: Wine quality data

The wine quality data set summarizes various properties of different wines. The variables include amongst others the amount of alcohol, the ph-value, the density, the residual sugar and a quality-score (ranging from 0-10) . There are two data files in the folder of this exercise, one for red wines and one for white wines. Load the data sets "winequality-red.csv" and "winequality-white.csv" using the `read.csv('winequality-red.csv', stringsAsFactors=TRUE)` command (same for `winequality-white.csv`).

a) Since the red and white wine data have the same variables, we wish to combine them into one big data frame. First, add a new column (`colour`) to both data sets taking the value `"red"` for red wines and `"white"` for white wines. Afterwards, bind the two sets together into one big data set. (**Hint**: `rbind()`)

b) Check the structure of the newly created data set. Turn the `colour` variable into a factor.

c) We wish to represent the wine data in two dimensions. Run a PCA and plot the two first PCs. Exclude the colour factor from the PCA (only numerical variables allowed). Also exclude the `quality` variable from the PCA. (**Hint**: `prcomp(..., scale.=TRUE)`)

d) Plot the PCs again with the colour of the points representing white and red wines. What can you observe?

e) We now want to see how a k-means algorithm would cluster the data represented in the 2D PCA-coordinates. Apply k-means clustering to the coordinates of the two first PCs with k=2 clusters. Set the random seed to 111 (`set.seed(111)`). Plot the created clusters. (**Hint**: `kmeans()`, `pca_obj$x`)

f) Apply again k-means clustering with k=2 clusters to the data, but this time set the random seed to 12 (`set.seed(12)`). How do you explain the result?

g) Use k=3 clusters and plot the results.

h) What would you say is the best number of clusters for this data? Answer based on your visual impression of the results and also by looking at the within-cluster sums of squares.

i) Apply hierarchical clustering to the coordinates of the first two PCs (choose the `ward.D2` method). Plot the created dendrogram. Plot the PCs again and indicate by colour the two biggest clusters of the dendrogram. (**Hint**: `hclust()`, `cutree()`)