Clustering using the kmeans algorithm

Exercise 1 Partition and matrix

Consider the Iris data set. Write a R code wich produces the partition matrix. Compute the gravity centers of the quantitative variables in the three classes using a matrix formula.

Exercise 2 The bell number

1. Show that the number of partition of n objects verifies

$$B_{n+1} = \sum_{k=0}^{n} C_k^n B_k$$

2. Compute manually the bell number for 1,2,3,4,5,6 objects.
3. Write a R program which computes the Bell number for n objects.

Exercise 3 Between-Within Variance relation

Consider n points from $\mathbb{R}^p$ with a partition into K classes of size $n_1, ..., n_k$. Let us note $\hat{\boldsymbol{\mu}}_k$ the gravity center of class k and $\hat{\boldsymbol{\mu}}$ the gravity center of the entire cloud of points.
Show that

$$\sum_k \sum_{i \in k} \|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k\|^2 + \sum_k n_k \|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}\|^2 = \sum_i \|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\|^2$$

Exercise 4 Clustering of the `crabs` (library MASS)

1. Load the `crabs` dataset form library `MASS`.
2. Plot the dateset using pairs() with a color for each specy and a different symbol per sex.
3. Cluster the dataset reduced to its quantitative variables into four cluster using the kmeans.
4. Run the algorithm with 1000 different initialization and keep track of the within sum of squares.
5. Comment the result.
6. Divide all quantitative variable by the most correlated variable to produce a new dataset.
7. Compare the partitions obtained using the kmeans with the 'natural' partition. Comment.
8. Try to cluster the data in 1 to 20 groups. Plot the within sum of squares in function of the number of clusters. Comment the figure.

Mixture Models

You need to install the package `mclust` *for the following exercises.*

Exercise 1 One dimensional mixture of Gaussians

1. Simulate a sample of size 1000 of a one dimensional mixture of two gaussians with respective means, variances and proportions $\mu_1 = 0$, $\mu_2 = 4$ $\sigma_1 = 1$, $\sigma_2 = \frac{1}{2}$, $\pi_1 = \frac{1}{3}$.
2. Use the kmeans algorithm to find two clusters.
3. From the kmeans output (classification) estimate the parameters of the mixture.
4. Use the `Mclust` function of the `mlust` package to estimate the parameters of the mixture:
   - try `Mclust` with modelNames="E"
   - try `Mclust` with modelNames="V"
5. Comment the diffrences between the three previons estimations.

Exercise 2 Bi-dimensional mixture

1. Load the dataset `faithful` (from the `mclust` library.
2. Plot and describe the data.
3. Run `Mclust` on the data and decribe the result:
   - The number of cluster.
   - The parameters (variance matrices, means vectors and proportions.
   - The classification.
4. Plot the output of the `Mclust` procedure and describe each of the 4 plots.
5. Run the `hclust` on the data using the Ward Criterion and compare the clustering of `hclust` and the clustering of `Mclust` for two clusters.
6. Run the `hclust` on the data using the Ward Criterion and compare the clustering of `hclust` and the clustering of `Mclust` for three clusters.
7. Comment on the results of the two previous questions.