Unsupervised Learning

The task of this exercise is to analyze voting data from the 111th US House of Representatives in 2009-2011. Votes in which names and positions are recorded are called “roll calls”. The file `rollcall-votes.Rdata` contains whether the 445 members of the House voted “yea” = 1 or “nea” = -1 in 1647 different voting decisions. If the vote = 0, the politician missed the voting. The file `rollcall-members.Rdata` contains further information about the members of the House such as party affiliation (Republican or Democrat) and which US state the politician represents. The data set comes from the website `www.voteview.com`.

Download the data sets `rollcall-votes.Rdata` and `rollcall-members.Rdata`. Load the data into R.

Exercise 1: Data Description and Visualization

1. How many Democrats and Republicans are in the House? (Report the special case Democrat-Republican separately.) Who has the majority?
2. Generate a variable for the number of votes each politician missed. How many politicians voted in all votings? Plot a histogram of the shares of missed votings with 100 bins.
3. Generate variables for the number of times each politician voted “yea” and “nea”. Make a scatter plot of the number of “nea” and “yea”. Use different colors to differentiate the points in the scatter plot by the party affiliation. Can we claim based on these results that the party could be a good predictor for the voting behavior?

Exercise 2: Principal Component Analysis (PCA)

1. Run a principal component analysis on the `votes` data set. How many principal components are there?
2. Calculate the proportion of variance explained by each principal component and plot the proportions for the first 10 principal components.
3. Plot the first two principal components and use color to differentiate the observations by the party affiliation.
4. With the help of the first principal component find politicians on the far right (very conservative) and far left (very liberal).
5. Find the votings which have the most extreme loadings for the second principal component. Analyze the voting behavior in these votings and come up with the interpretation of the second component.

Exercise 3: k-means Clustering

1. Run a k-means clustering procedure to detect 2 clusters in the data. How many Democrats and Republicans are in each cluster?

2. Run a k-means clustering procedure for $k \in \{2, \ldots, 20\}$ and plot the within cluster sum of squared errors. Decide based on the plot what the optimal number of clusters is.

3. Take the graph with two principal components and use color to visualize the clusters you found in the previous task.

4. Run a k-means clustering procedure to detect 6 clusters with `nstart = 1` and `nstart = 20`. Print out the within cluster sum of squared errors for both `nstart` values. What can you say about the importance of `nstart` on the convergence to the global optimum?