K-Means Clustering in R – Exercises

K-means is efficient, and perhaps, the most popular clustering method. It is a way for finding natural groups in otherwise *unlabeled* data. You specify the number of clusters you want defined and the algorithm minimizes the total within-cluster variance.

In this exercise, we will play around with the base R inbuilt k-means function on some *labeled* data.

Exercise 1

Feed the columns with sepal measurements in the inbuilt iris data-set to the k-means; save the cluster vector of each observation. Use 3 centers and set the random seed to 1 before.

Exercise 2

Check the proportions of each species by cluster.

Exercise 3

Make a plot with sepal length on the horizontal axis and width on the vertical axis. Find a way to visualize both the actual species and the cluster the algorithm is categorized into.

Exercise 4

Repeat the clustering from step one, but include petal measurements also. Does the clustering reflect the actual species better now?

Exercise 5

Create a new data-set identical to iris, but multiply the “Petal.Width” by 2. Are the results different now?

Exercise 6

Standardize your new data-set so that each variable has a mean of 0 and a variance of 1; check once more if multiplying by two makes a difference.

Exercise 7

Read in the Titanic train.csv data-set from Kaggle.com (you might have to sign up first). Turn the sex column into a dummy variable, =1, if it is male, "0" otherwise, and "Pclass" into a dummy variable for the most common class 3. Using four columns, Sex, SibSp, Parch and Fare, apply the k-means algorithm to get 4 clusters and use nstart=20. Remember to set the seed to 1 so the results are comparable. Note that these four variables have no special meaning to the problem and dummy data in k-means is probably not a good idea in general – we are just playing around.

Exercise 8

Now, how would you describe, in words, each of the clusters and what the survival rates were?

Exercise 9

Now, 4 clusters was an arbitrary choice. Apply k-means clustering using nstart=20, but using k from 2 to 20 and store the result.

Exercise 10

Plot the percent of variance explained by the clusters (between sums of squares over the total sum of squares). What seems like a reasonable number of clusters, according to the elbow method?