Exercise:
K Nearest Neighbor

Machine Learning and Prediction Modelling

Exercise 1: Crabs data set

The `crabs` data set describes different morphological measurements of two different crab-species (see https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/crabs.html for a detailed description of the data). We want to train a KNN-classifier on the crabs data set.

a) The data set is contained in the R package `MASS`. Load the package in order to access the data frame `crabs`.

b) Get a first impression of the data and make sure that everything is coded correctly. Create a pairs plot to get a visual impression.

c) Fit a knn classifier to the data using `k=3`. We try to model the species (`sp`) of the crabs based on the five morphological measurements (`FL`, `RW`, `CL`, `CW`, `BD`). (**Hint**: `knn()`, you need to load the package `class` for the `knn()` function)

d) Display the confusion matrix of the predictions (on the training data itself) and calculate the training error.

Exercise 2: Cardiotocography data set

The Cardiotocography data set is a collection of fetal cardiotocograms and the associated diagnostic features. Each cardiotocogram was also classified by three expert obstetricians with regard to the fetal state (normal, suspect or pathologic). See https://archive.ics.uci.edu/ml/datasets/Cardiotocography for more information. We work with a reduced version of the data set, which only includes a sample of the total records.

a) This data set is supplied as a .csv file. Read in the data with the command `fet <- read.csv('Cardiotocography.csv', stringsAsFactors=TRUE)` and get a first impression of the data. There should only be one factor in the data which is the status of the fetus. How many rows does the data contain? How many observations are there with `normal`, `suspect` and `pathologic` status, respectively?

b) We want to compare the training and test error when using k-nearest neighbor classification on the cardiotocography data. The goal is to predict the status of a fetus based on the available measurements. Remove a random sample from the data (size=200), we will use this subsample as a test data set. To take a random sample from a data frame, we can make use of the `sample` function. The `sample`

function can be given a vector as an argument, and it will take a random sample of a specified size from this vector, for example:

```
sample(1:10, size = 4)
```

```
## [1]  8  4  9 10
```

Try to use the `sample` function to extract 200 random rows from our data. The remaining data will be used as training data. Apply KNN to the training data with k=4 and calculate the training and test error.

c) We now want to compare the training and test error for different k. Write a loop in which you fit a knn classifier to the training data with k ranging from 1 to 30. Collect for each k the training error and the test error in a table.

d) Plot the test and training error rates against k. Which k should be chosen based on this quick analysis? Why is this the best k?