K Nearest Neighbors (KNN)

Assignment 02 – report by Frédéric Charon

In this assignment we implement the code for the K nearest neighbor algorithm. The goal is to first implement some trainings data and then use that data to classify a new data point using the KNN method.

KNN works as followed: we get a new data point, look at its closest other points and choose the most common class for our new point. The value k determines at how many closest points we look.

For our code we initially import a dataset containing different examples of plants for the training data, these plants are divided into three different classes (Iris-setosa, Iris-versicolor and Iris-virginica) and classified by four different features (petal length, petal width, sepal length and sepal width).

As shown in the assignment our program is made of three different parts:

1. Calculate the distance using the Euclidean distance function
2. Find the k nearest neighbors
3. Make the prediction for the new data point

The first part is simply an implementation of the Euclidean distance function which defines the distance between two data points, in our case using the four different features as N.

$$d(p,q) = \sqrt{\sum_{i=1}^{N}(q_i - p_i)^2}$$

After that we write the k nearest neighbors method. Here we have two inputs: a new data point and a value for k. Now we measure the distance between each point of our dataset and the new data point. We compare them and pick the k closest data points.

In the last step we look at these k points we just picked and what class they belong to. The most common class for these points is going to be the class we choose for our new data point.