Introduction to Data Science

# K-Nearest Neighbors

## Assignment 2

*Submitted by*

Andras Csepreghy

xgj708

28th of February, 2019

UNIVERSITY OF
COPENHAGEN

# Introduction

This assignment includes basic classification with K-Nearest Neighbors classifier classifier, a non-linear, non-parametric method for classification. Then applying cross-validation for model selection and standard data normalization for preprocessing.

# Exercise 1

In the first exercise I used sci-kit learn's nearest neighbors classifier with the predefined k of 1. The splitting of training and test sets were predefined by the dataset.

Results: Results show that the accuracy obtained even with $k = 1$ is surprisingly good, 0.946. These good results may be explained by the quality of the data collected.

# Exercise 2

In this exercise I had to decide a reasonably good value for $k$ by calculating the accuracy of the classifier for every given k. For personal experimentation I extended the number of $k$s to test.

Results:

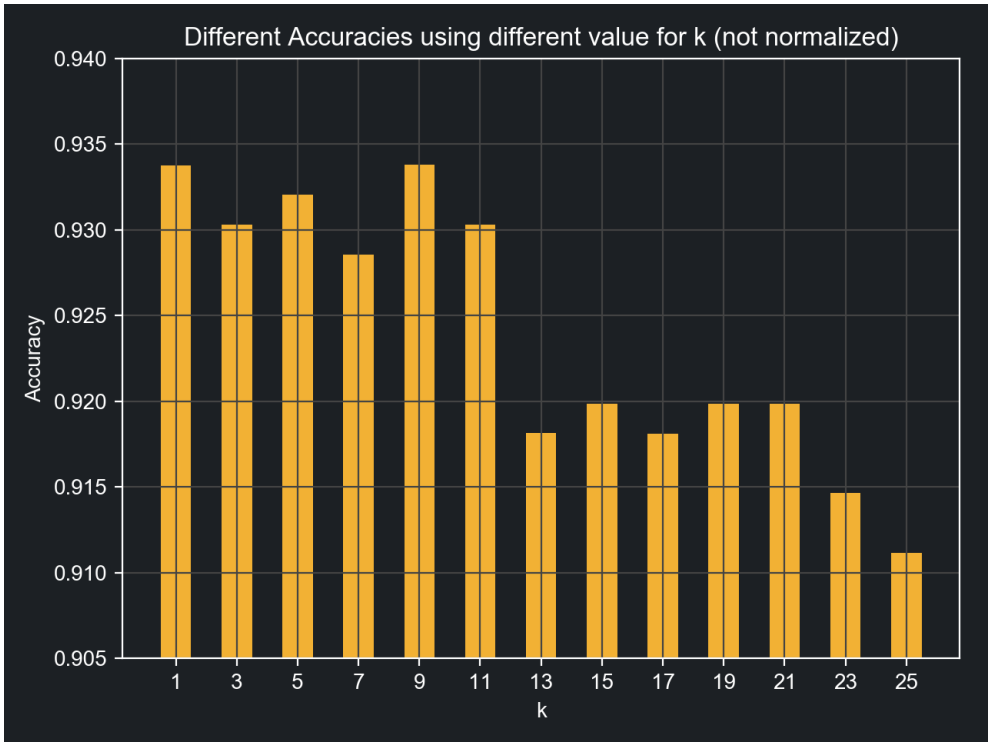


**Fig 1**: Different accuracies using different k values

# Exercise 3

Results: Without normalization the best value for k turned out to be 9, only a little more accurate than k = 1 with accuracy being 0.9408. Even though this looks lower than the accuracy of 1-NN, however k = 9 performs better when looking at the mean of accuracies in cross-validation.

# Exercise 4

There were 3 different approaches presented for normalization out of which Version 1 is the correct one. The easy, cheating way if concluding this could be that this is the one that sci-kit learn uses in their documentation and they better know how to normalize data. But I'll provide a more thought out answer too.

Version 2 is incorrect because it normalizes the data twice, both for training and testing data, which gives different variances for the two datasets. Normalization should only happen to the dataset on which we train our algorithm.

Version 3 is incorrect, because it combines the two dataset for normalizing after which the algorithm would be trained using the normalization obtained by using test data. We shouldn't have access to the test set while preparing and training any machine learning algorithm and only use test set for testing.