Sabancı Üniversitesi

Machine Learning (CS412)

Assignment one

Student name: Hagverdi Ibrahimli

Student id: 00030014

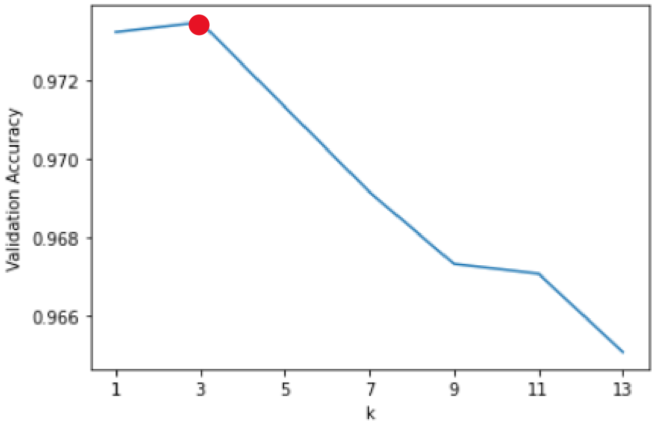
Submission title: Detailed report for the Machine Learning course (CS412), assignment one.

Link for the Colab notebook:
https://colab.research.google.com/drive/1Jm0J1tvRBBAFMllEc42iI_XiiRAQxwzG?usp=sharing

The aim of this report is to explain the steps taken to solve the first assignment of the Machine Learning course. The problem definiton is to implement a machine learning model which can predict (preferably with low error rate) a given sample from the MNIST dataset. For this purpose, we are going to train a k-NN classifier and optimize its performance by finding an optimal value for the nearest neighbors (I.e. k) parameter with the help of Scikit-learn library. First of all, MNIST dataset provides us with 60,000 training data and 10,000 test data. To find the optimal nearest neighbors parameter, we should split the training data into two portions: development data and validation data, with respective portions of 80% and 20%. It is noteworthy to mention that, we shuffle the training data beforehand splitting it to prevent our model to be biased towards any specific patterns or sequences in the data. Next, we train models on the development data with different values for the nearest neighbors parameter from the list: *[1, 3, 5, 7, 9, 11, 13]*. Then we measure each of their performance on the validation data so to choose the one with the highest accuracy. Below is the table attached depicting the accuracy rate of models trained with different values for the nearest neighbors parameter from the list aforementioned.

Nearest Neighbors parameter (K)	Accuracy of the model on the validation dataset (percentage)
1	0.97325
3	**0.9735**
5	0.9713
7	0.9691
9	0.9673
11	0.967
13	0.965

As highlighted on the table, the nearest neighbors parameter with the value of 3 showed the highest performance on the validation dataset with the accuracy of 97.35%. For this reason, we select the value 3 for the nearest neighbors parameter and train our new model on the full training dataset (i.e. development and validation data combined) which gives us the accuracy of 97.05% on the test data. Also, we made a graph to better visualize the increase in the nearest neighbors parameter and change in the model accuracy. The red dot points to the k value which has the highest accuracy.


Validation Accuracy
k