CS412 – HW1 REPORT – TAN UFUK CELIK – 28285

https://colab.research.google.com/drive/1P3jErd6hJMW6qQ-7JhS60BA8jZh8SrUn?usp=sharing

• The MNIST dataset is a dataset consisting of images of handwritten digits. This dataset is widely used to test the effectiveness of classification algorithms. Our problem in this assignment is to examine how the k-NN algorithm works with this dataset.

• There are a lot of data in the world and the types of these data can be different from each other, so it is very important to translate these data into a language that computers can understand and to clean the data and make it workable. In order to transform this data into clean data, we can do more operations such as adding, subtracting, averaging. After thoroughly understanding the data, we need to know which classification method to choose. Because the situation, type and number of data varies how well the classification methods will work. k-NN performs well in the MNIST dataset because classification of handwritten digits relies on being able to distinguish them from other digits with similar characteristics. k-NN can fit well with these features as it allows data to be classified using direct distance measures. We choose our classifier and we need to separate this clean data into development, validation and testing. In this way, we can find the number of accuracy in the training and estimation of the model and make the right decisions.

k | Validation Accuracy

K=1 | 0.103

K=3 | 0.1055

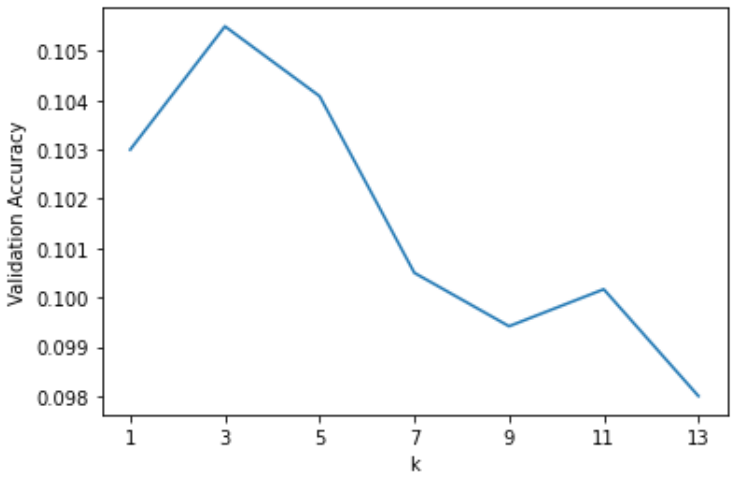
K=5 | 0.10408333333333333

K=7 | 0.1005

K=9 | 0.09941666666666667

K=11 | 0.10016666666666667

K=13 | 0.098



• As we can see, the best validation accuracy of 0.1055 was achieved with k=3. Therefore, I selected k=3 as my model. Also if you can print these informations in the code part.

• I obtained the best validation accuracy of 10.55% using the k-NN approach with k=3. Then, I evaluated this model on the test set and achieved an accuracy of 10.85%."

EXTRA WORK AND NOTES:

1-NOTE: The question that comes to my mind: Why do we mix only the training data and not the test data?
Anwer: When mixing data, it is usually sufficient to mix only the training data (X_train and Y_train). Test data (X_test and y_test) it is already a different dataset from the training data, so there is no need to shuffle them. Data may prevent us from properly assessing the model's real-world performance. Because the model have seen and learned the data. Since the test data is a completely different data set, how much of the model to this data It is important to keep test data separate to see if it fits well.

2-NOTE: With #random.seed() we can always do the same mixing by keeping the Seed value constant. It's already written at the very beginning of the code.

3-NOTE: The validation set is used to check the accuracy and generalizability of the model during training. Whether our model encounters overfitting or underfitting problems during training. The validation set is important to understand and fix these problems. on training data While the optimizations give good results for the training data of the model, the new data it encounters in real life performance is not always guaranteed. In this case, part of the training data by separating it as a validation set, we try to predict the performance of the model on data that it may encounter in real life.
