CS412 Homework 1 Report

Uğur ÖZTUNÇ 28176

Colab Notebook Link

https://colab.research.google.com/drive/1et1owBP3D1gPHSx8jKCTpuvZwBsdJI7p?usp=sharing

1. Problem

In this assignment, we were asked to develop a supervised learning model that would help to label handwritten digits. To solve this classification problem, an application that implements a k-NN classifier model was developed in Python supplemented with related libraries and a large MNIST dataset.

2. Dataset

To talk a little bit about this dataset, it consists of two sets, one with 60,000 samples for training the model and other with 10,000 samples for testing the model developed. Each sample represents a grayscale image of a handwritten digit (0-9) consisting of 28x28 pixels that can be seen in the figure below. These samples have a feature for representing the gray-level value between 0 and 255 of each pixel of the image.

Figure 1: Samples from the MNIST dataset

3. Preprocessing

Before starting the training stage, the dataset with 60,000 samples was shuffled first, then it was split into two separate sets: 80% (48,000 samples) reserved for training phase and the remaining 20% (12,000 samples) reserved for validation phase. No other preprocessing operations, such as feature extraction, were performed on dataset.

4. Training

In training phase, the classifier model was trained with the training set with 48,000 samples, with k values odd numbers from 1 to 13, since using even values for k might increase the chance of tie situations. After training, at each k value, remaining part of the set was used for validation phase of the classifier, then the predicted Y values of the classifier were compared with the true Y values of the validation set in order to calculate the accuracy score of the classifier for every k value. The accuracy results for each k value are shown below.

k value	Accuracy
1	97.217%
3	96.992%
5	96.925%
7	96.842%
9	96.683%
11	96.558%
13	96.392%

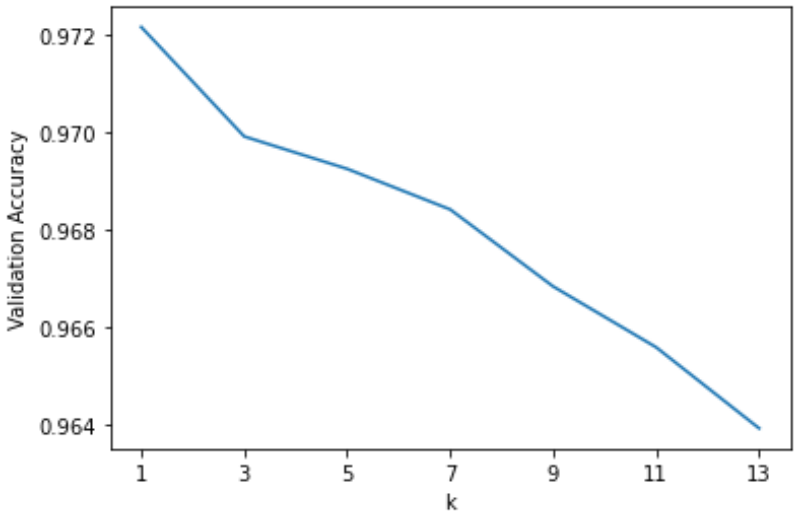

Figure 2
(Accuracies were rounded to 3-digit decimal.)

5. Testing

We have obtained the best results on the validation set with the k-NN approach using a value of 1 for k parameter. After determining the best value for the k parameter, a new classifier was trained with whole MNIST training dataset (all 60,000 samples). Then the classifier was tested with the testing dataset, which contains 10,000 samples, that is completely discrete from the training set. The final performance of the model with chosen k value of 1 was measured as 96.91% accuracy. In conclusion, it can be said that the final model can be considered as a consistent model, since the performance obtained at a distinct test set shows parallelism with the validation accuracy on training set.