Emre ZEYTINOGLU

28190

CS412 First Homework

Google Colab Link:

https://colab.research.google.com/drive/1iFcDp3Vyb3ymPGxjRKfiUVEoTBHIfCYy?usp=sharing

Our problem is to find best approach to MNIST dataset. Data is reached with the help of the Keras Library and then data is fixed to reach sizes that are necessary in order to continue array process.

The shuffle() function is used to randomly shuffle the training data X_train and its corresponding labels y_train. This is often done to ensure that the data is evenly distributed and to prevent any bias that may be introduced during the data collection or preprocessing stages.

After shuffling the data, the train_test_split() function is used to split the data into training and testing sets. In this case, 20% of the shuffled data is randomly selected and set aside as the test set, while the remaining 80% is used as the training set. The test set is used to evaluate the performance of the model after it has been trained on the training set.

It is important to split the data into training, validation, and testing sets to ensure that the model is able to generalize well to new, unseen data. The training set is used to train the model, while the validation set is used to tune the hyperparameters of the model and

prevent overfitting. The testing set is used to evaluate the performance of the model on new, unseen data.

The size of the training, validation, and testing sets can vary depending on the size of the dataset and the complexity of the model. In this case, 20% of the data was set aside as the testing set, which is a common split ratio. The remaining 80% of the data was split into the training and validation sets, which can also vary depending on the specific use case and available resources.

Based on the validation accuracy results, the best performing approach for this model is the k-Nearest Neighbors (k-NN) algorithm with k=1.

Here is a table summarizing the validation accuracies for different values of k:

k value	**Validation Accuracy**
1	0.9712
3	0.9708
5	0.9699
7	0.9677
9	0.9663
11	0.9655
13	0.9648

As the best validation accuracy of 0.9712 was achieved with k=1, this was chosen as the model for the test data.

The test results for this model are as follows: "We have obtained the best results on the validation set with the k-Nearest Neighbors (k-NN) algorithm using a value of k=1. The result of this model on the test data is 100% accuracy."