Cs412 Machine Learning Homework 1 Report -Zeynep Melis Meriç 24170

Problem: MNIST dataset consists of handwritten numbers, the problem is to predict the number in the set.
We worked with Decision Tree Classifier and K-nearest neighbours classifier and decide which one to use according to their accuracy score.
Train Set: Train set is the actual dataset to train the model for performing various actions like training the classifier. Validation and training sets are available during training the classifiers.
Validation Set: Usually %20 or % 30 of the training set, in the project I worked with %30 of training set as validation. Validation set is used to tune the parameters
Test Set: Test set is used to test the performance of the classifier.
After downloading the dataset we take portion(10) of the data to work faster. Then we shuffle the train data and train labels with random_state 20. Hence we get the random mix of the data. Then we split the data with sklearn.model_selection function split, after spliting we got 80% of train(development) and 20% of validation data to work with. With development portion of the training data we trained decision tree and k-nn classifiers to select the best one. In the table the validation accuracies for different approaches and parameters can be seen;

Classifier and Parameter	Validation Accuracy
K-nn k=1	95.50000%
K-nn k=3	95.16667%
K-nn k=5	94.33333%
Decision Tree Classifier m=1	39.41667%
Decision Tree Classifier m=5	39.41667%

m= min_samples_split
From the table, we could decide that the best classifier for our case is K-nn with k=1. Then we tested classifiers with the test dataset and the conclusion was the same with the output:

```
Testing Accuracy with decision tree classifier = 38.10000%
Testing Accuracy with K-NN= 88.40000%
```

We have obtained the best results with the k-nn classifier (parameters= k= 1.) , giving a digit classification accuracy of 88.40000% on test data. In the output we could see that the predictions are the same as the true labels. In the confusion matrix we could see that the true labels are mostly match the prediction labels.

Confusion matrix:

```
[[ 84   0   0   0   0   0   1   0   0   0]
 [  0 126   0   0   0   0   0   0   0   0]
 [  1   1 103   1   0   0   2   3   5   0]
 [  0   1   0 102   0   0   0   2   1   1]
 [  1   1   0   0 100   0   1   0   0   7]
 [  1   0   0   2   0  76   2   0   4   2]
 [  3   0   0   0   1   0  83   0   0   0]
 [  0   3   0   0   1   1   0  94   0   0]
 [  3   0   1   2   1   2   1   1  76   2]
 [  0   1   0   0   0   0   0   0   3  90]]
```

Link: https://colab.research.google.com/drive/1V7-pjbungNp6NZxMjXInFF9yLPuBXWo_