Tutorial 1 - Intro to Machine Learning

Ioannis Manousaridis

Task 1

Table 1.

Training set	Testing test	Accuracy Score
80%	20%	0.8974
50%	50%	0.8205
20%	80%	0.8205
10%	90%	0.32522

Conclusions:

It is obvious from the scores that the last splitting, the one that used 50% for training and 50% for testing, provided the best results. Although, it is not very cleared if the score of accuracy decreases as the samples for training are decreasing. This is visible from the a and b example. The a example uses less data for training but it has better results than the b example. In general, though, it make sense that the more data for training, the better results in tests as it can been seen by the example c. This is also supported by this example in which the 50-50 splitting provided worst results from the splitting 80% training dataset and 20% testing dataset. The last gave a result of 0.8974. Summarizing, more tests are required to make better and more confident conclusions.

Task 2

Changing the augmentation scale and ratio in images produced the results in table 2:

Table 2.

Scale	Ratio	Accuracy Score
4	0.01	0.8571
6	0.05	0.8163
4	0.1	0.8573

Comments & conclusions:

The original dataset for test size 20% gave a result of 0.8974. Using the technique of data augmentation the results are supposed to improve because the size of data for training is increased. In this case, this didn't happen and the results weren't improved. The best result with data augmentation technique is 0.8573 which is 4% worse than the one with the original dataset.

Task 3

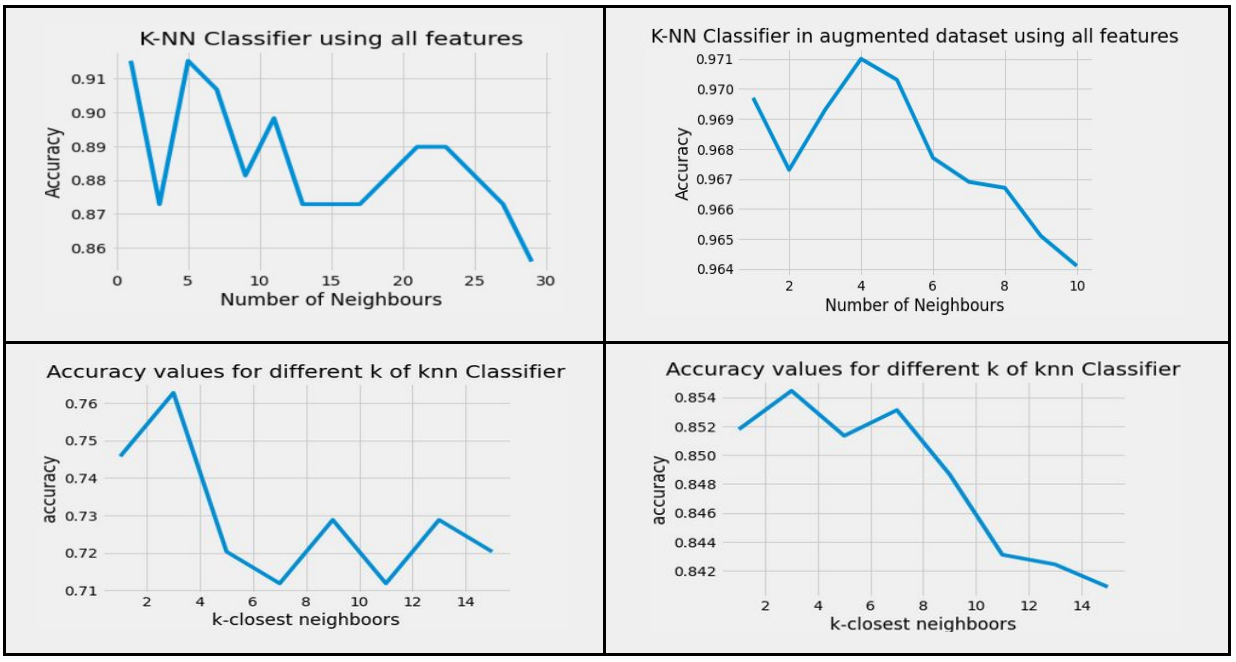

***Figure 1**: On the plots are the K-NN classification accuracies for different values of k. On the upper plots the mnist dataset was used. The plot on the left is without augmentation by using all features and on the right is with augmentation and by using all features. On the down plots the accuracies about the fminst dataset are displayed. One the left is without augmentation and all features and on the right is with the augmented data and with all features.*

Conclusions

It is obvious from Figure 1 that the data augmentation improved the results in both datasets. The accuracy range of the mnist dataset is 0.85-0.91 and with the augmented data it is 0.964-0.971. The results are highly improved and provide an excellent result. The same applies for the fmnist dataset. The results are slightly lower and this is normal because the fmnist dataset is more complicate. The initial range is 0.71-0.77 and with augmented data it is 0.84-0.855. In general, the data augmentation technique improved the results of the K-NN classification by a 10% and also reduced the length of the range provided more stable accuracies. This is result was expected because with the augmented data the classifier had more data for training.