Tutorial 2 - Ioannis Manousaridis

Task 1

Train

k \ m	N	5	2
1	1	1	1
3	0.9874	0.9572	0.9372
5	0.9974	0.9572	0.9397

Test

k \ m	N	5	2
1	0.9415	0.9181	0.8187
3	0.9473	0.9298	0.8888
5	0.9473	0.9298	0.8888

Train

k \ m	N	3	2
1	0.9904	1	0.9904
3	0.9714	0.9523	0.9714
5	0.9619	0.9524	0.9524

Test

k \ m	N	3	2
1	0.9777	1	0.9777
3	1	0.9777	1
5	1	1	1

Figure 1: K-NN classification with m-best features. Up for the wdbc dataset and down for the iris dataset. On the left side the training scores and on the right the testing.

Comments:

In this task, two datasets, the iris and the wdbc, were used for k-nn classification without PCA. Different combination of k neighbours and m-best features were selected. The accuracies as shown in the Figure 1 are all very high.

Task 2

Train

k \ m	N	5	2
1	1	1	1
3	0.9874	0.9874	0.9572
5	0.9773	0.9748	0.9472

Test

k \ m	N	5	2
1	0.9415	0.9473	0.9239
3	0.9473	0.9415	0.9239
5	0.9473	0.9415	0.9122

Train

k \ m	N	3	2
1	1	1	1
3	0.952	0.942	0.962
5	0.961	0.942	0.952

Test

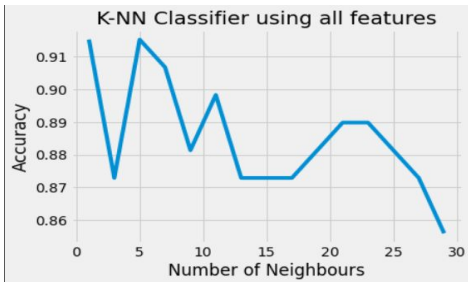
k \ m	N	3	2
1	0.977	0.933	0.933
3	0.977	0.977	0.889
5	0.977	1	0.977

Figure 2: K-NN classification with m-best features after PCA. Up for the wdbc dataset and down for the iris dataset. On the left side the training scores and on the right the testing.

Comments:

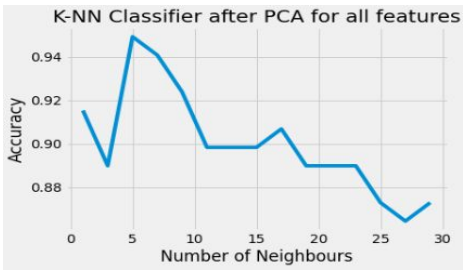
Task 2 is similar to task 1 with the difference that the PCA procedure was applied to both datasets. Comparing the figure 1 and 2, we notice that the scores for the wdbc have been slightly improved, whereas the iris' results are quite similar.

Tasks c-g for mnist dataset



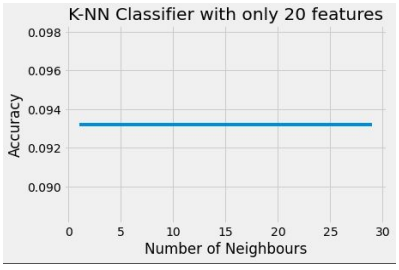
k=1 achieved highest accuracy of 91.53%

(a)



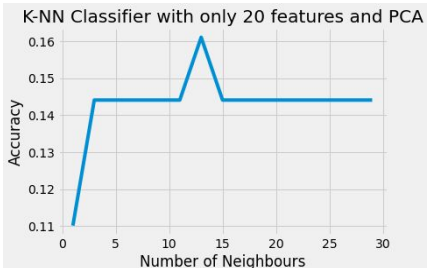
k=5 achieved highest accuracy of 94.92%

(b)



k=1 achieved highest accuracy of 9.32%

(c)



k=13 achieved highest accuracy of 16.10%

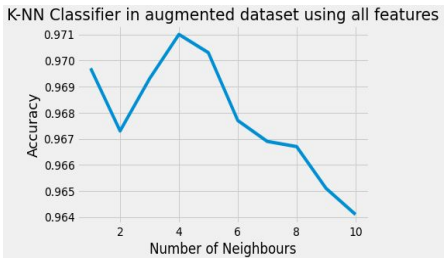
(d)

Comments:

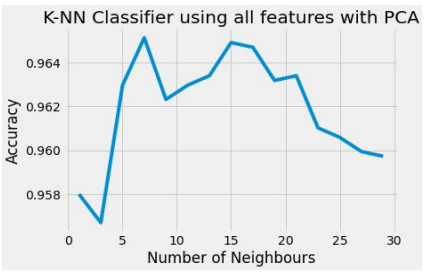
In Figures 3a, 3b, there are the results of the K-NN classification on the mnist dataset using all features without and with PCA accordingly. The PCA improved the accuracy by 3%.

Comments:

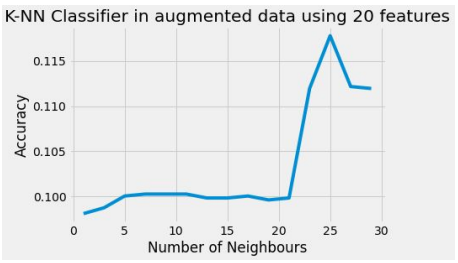
In Figures 3c, 3d, there are the results of the K-NN classification on the mnist dataset but by using only 20 features without and with PCA accordingly. The accuracies are extremely low in each case.



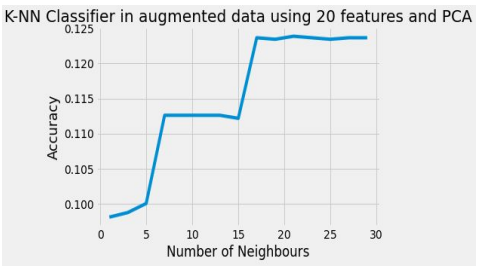
k=4 achieved highest accuracy of 97.10%
(e)



k=7 achieved highest accuracy of 96.51%
(f)



k=25 achieved highest accuracy of 11.78%
(g)



k=21 achieved highest accuracy of 12.39%
(h)

The data augmentation increased the dataset from 1568 samples to 61568. The K-NN classifications with the augmented data can be seen in the e,f pictures. Comparing the a,b and e,f plots, it is obvious that the data augmentation improved the accuracy results.

In the plots g,h are the results of the augmented dataset with only 20 best features. As we can see the results are still extremely low.

Conclusions: Both the PCA and the data augmentation improve the results. The data augmentation provided better results than the PCA procedure. The worst results were noticed when only 20 features were used and neither the PCA or data augmentation improved them. Finally, in the figure 4, we can see the importance of each eigenvector, from the most to the least in important.

Figure 3: K-NN classification plots for different cases in the mnist dataset.

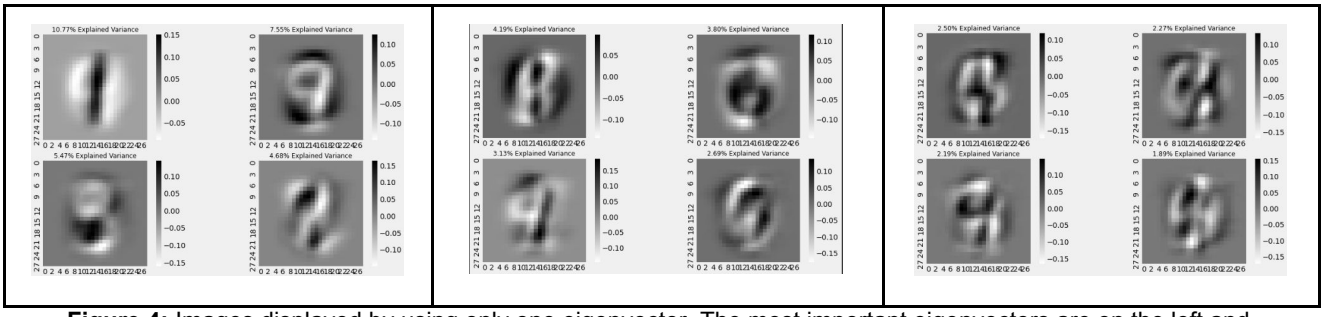
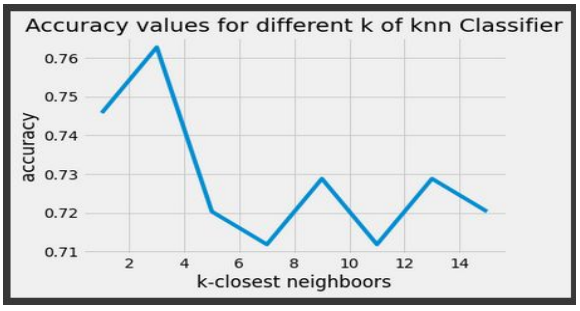


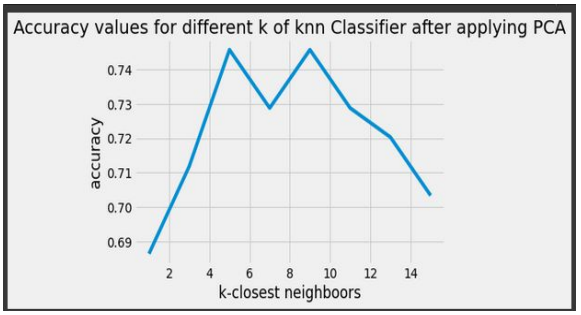
Figure 4: Images displayed by using only one eigenvector. The most important eigenvectors are on the left and the importance decreases on the right.

Tasks c-g for fashion mnist dataset

Below are the similar plots for the for fashion mnist dataset.

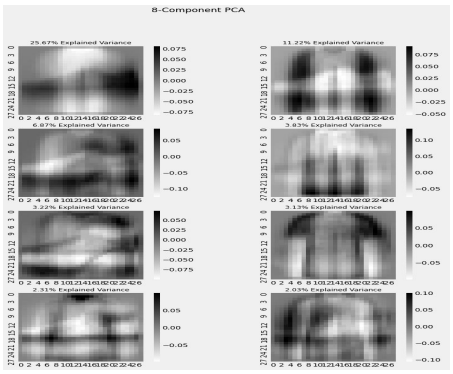


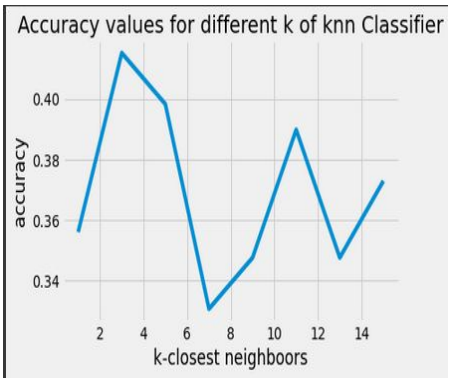
k=3 achieved highest accuracy of 76.2%



k=5 achieved highest accuracy of 74.5%

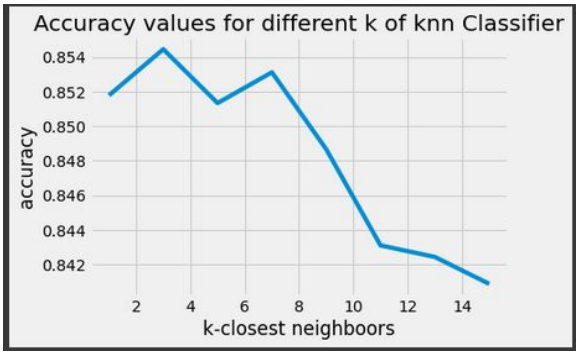
***Figure 5**: On the left are the classification results without using PCA and on the right are classification results after applying PCA for all the features.*



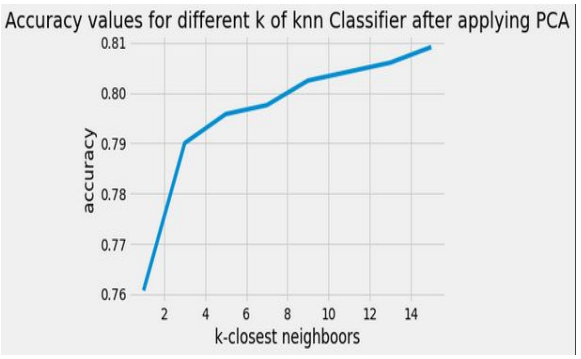


k=3 achieved highest accuracy of 41.53%

***Figure 6**: Eigenvectors corresponding to the greatest eigenvalues on the left and classification results for only 20 best features on the right.*

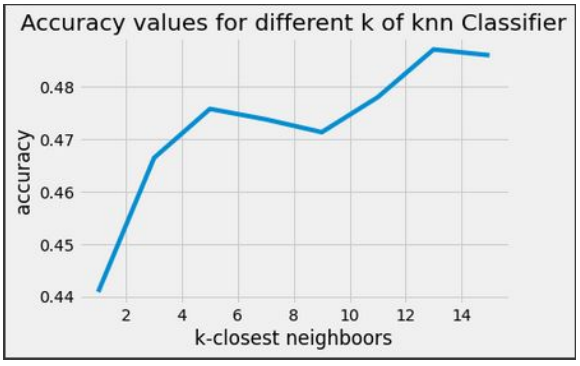


k=3 achieved highest accuracy of 85.44%

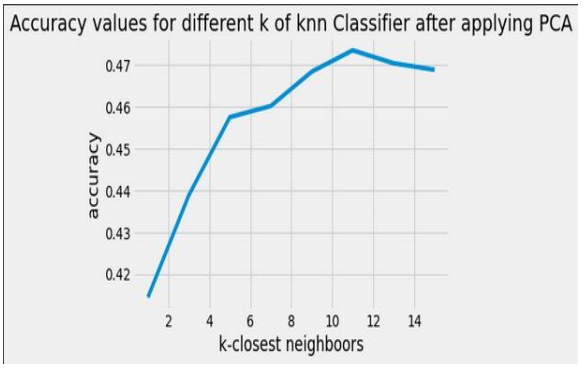


k=5 achieved highest accuracy of 80.91%

***Figure 7**: On the left classification results before applying PCA and on the right classification results after applying PCA, when using all features and data augmentation.*



k=13 achieved highest accuracy of 48.7%



k=11 achieved highest accuracy of 47.36%

***Figure 8**: On the left classification results before applying PCA and on the right classification results after applying PCA, when using the 20 best features and data augmentation.*

Conclusions:

The initial classification at the original dataset with all features provided good results around 75%. The PCA procedure improved the results in general but didn't achieve the highest accuracy. As it was expected by using only the 20 best features the results were quite bad. The data augmentation technique improved quite much the results even with the 20 best features.

In general, the data augmentation improves more the accuracy than the PCA, but both methods help in the improvement of accuracy.