UUN:S1525701

Task2

Task 2.3

Determinants of covariance matrix of:

- Class 1: $8.94933712801e^{-25}$
- Class 2: $2.82852645141e^{-37}$
- Class 3: $2.52810991847e^{-65}$
- Class 4: $2.4690437482e^{-23}$
- Class 5: $6.40249199698e^{-44}$

Confusion matrix for the test data reduced to 100 dimensions:

Predicted	1	2	3	4	5
Actual					
1	146	13	4	33	4
2	11	130	13	39	7
3	6	3	156	18	17
4	24	22	3	143	8
5	8	8	22	13	149

After we normalize the table we get the following classification rates:

- Class 1: 74.9%
- Class 2: 73.9%
- Class 3: 78.8%
- Class 4: 58.1%
- Class 5: 80.5%

Task 2.4

For the investigation on the effect of dimensionality on classification rate I used a Python script *classification_performance.py* , which can be found in *Task2* directory.

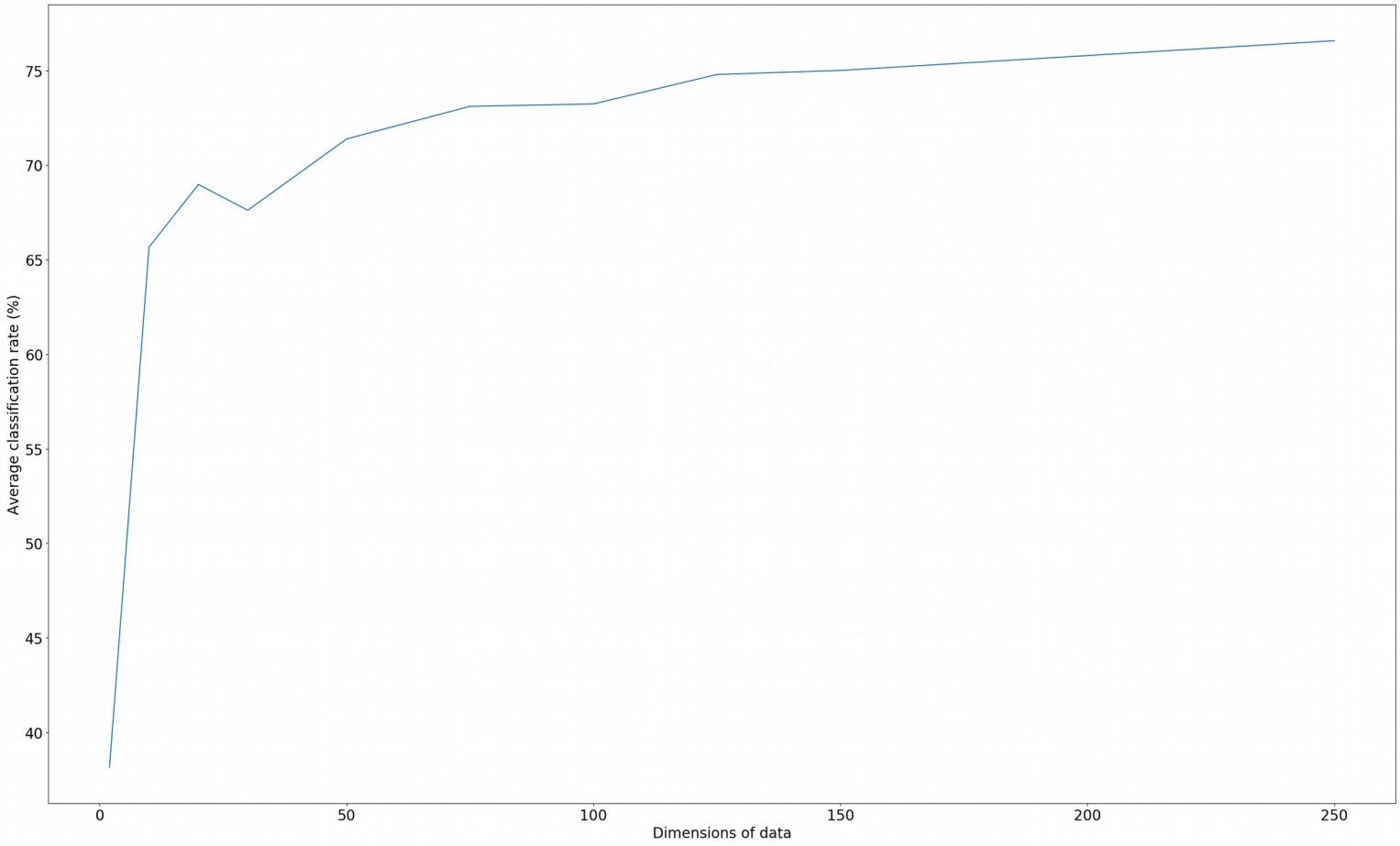
Files related to this task:

- *classification_performance.py – script to perform the experiment and plot the results*
- *performance_results.txt – full results of classification performance with different dimensionality*

The experiment

We train the model with train_x data reduced to a specified dimensionality and test the model on the reduced dimensionality test_x data. Additionally, the data is centred around the mean of the train data.

The experiment has been run with the following dimensions: 2, 10, 20, 30, 50, 75, 100, 125, 150, 250. A graph with the average classification rate is presented below.



Results

Average classification performance increases proportionally with the increase of dimensionality. However, for much higher dimensional data we can observe the "curse of dimensionality", where data is too sparse to classify efficiently and performance drops (not shown on the graph). This is the reason we apply Principle Component Analysis to the data to reduce its dimensions.

To aid in spotting the general trend in classification rate, below are listed exact results for 2, 100, 250 dimensional data. We can see that for the range [100,250] dimensions classification rates increase by a much lesser margin than for the range [2,100].
Full results can be found in *performance_results.txt* in *Task2* directory.

Classification rate for data with dimensionality reduced to: 2

- Class 1: 37.9%
- Class 2: 40.8%
- Class 3: 28.6%
- Class 4: 42.5%
- Class 5: 41.0%

Average classification rate: 38.16%

Classification rate for data with dimensionality reduced to: 100

- Class 1: 74.9%
- Class 2: 73.9%
- Class 3: 78.8%
- Class 4: 58.1%
- Class 5: 80.5%

Average classification rate: 73.24%

Classification rate for data with dimensionality reduced to: 250

- Class 1: 76.1%
- Class 2: 80.0%
- Class 3: 88.5%
- Class 4: 59.5%
- Class 5: 78.8%

Average classification rate: 76.58%