Exercise 2a report

The script running exercise 2a uses the scikit-learn library version 0.22.1 and the numpy library version 1.18.1 and methods from another file called dataPreparation, where the methods for loading a CSV file and for deleting useless data parts are written.

For the RBF kernel, the SVC class and the model_selection class of the scikit library are used to compute, for each parameter combination, the accuracy of the 10 folds (10-fold cross-validation). The means of those accuracies are used to find the best parameters. Of course, it is very hard to find the two perfect parameters: the optimal C for a certain gamma won't necessary be the best C for another gamma. To find good parameters, a grid search method is used, with 5 different values for both parameters.

For the linear kernel, instead of the SVC class, the LinearSVC class (always from the scikit library) is used, since it is supposed to be more efficient than SVC with linear kernel. A range is given (the results here are in the range 0.1 and 30) and an algorithm calculates the accuracy (using 10-fold CV) of different values of C between this range. After some results between this range, the algorithm zooms in a new range, located between the C with the best accuracy and one of its two neighbours, the one with the best accuracy. Everything starts again within this new range. It stops after a certain number of iterations.

We can see that for the linear kernel, the accuracy varies a lot with C. It can jump by 2% with very small differences of C, and there are some precision peaks in different places, making it difficult to find the perfect optimal C.

The RBF kernel has a better accuracy, more than 90% with every parameter combination tested here. With a gamma value of 2.1004e-07, the accuracy is higher for every C tested, and for C = 4, the accuracy is higher for every gamma tested. There is probably at least a local maximum accuracy near to C=4 and near to gamma = 2.1004e-07.

Results:

RBF kernel

Average accuracy during cross-validation for all parameter values:

C	Gamma	Accuracy
0.25	0.25 * auto*	94.04%
	0.5 * auto	95.47%
	1 * auto	96.72%
	2 * auto	97.34%
	4 * auto	96.95%
0.5	0.25 * auto*	94.81%
	0.5 * auto	96.28%
	1 * auto	97.33%
	2 * auto	97.90%
	4 * auto	97.71%
1	0.25 * auto*	95.64%
	0.5 * auto	96.94%
	1 * auto	97.80%
	2 * auto	98.25%
	4 * auto	98.09%
2	0.25 * auto*	96.33%
	0.5 * auto	97.44%
	1 * auto	98.11%
	2 * auto	98.42%
	4 * auto	98.19%
4	0.25 * auto*	96.93%
	0.5 * auto	97.75%
	1 * auto	98.23%
	2 * auto	98.45%
	4 * auto	98.19%

*auto is the auto value from the SVC class from the sklearn.svm library (Sci-kit learn). The auto gamma is here: 2.1004e-07

Average accuracy on test set with optimized parameter values:

Parameters:

10-fold CV

C = 4

gamma = 4.2009e-07

Accuracy**: 98.55%**

Linear kernel

Average accuracy during cross-validation for all investigated kernels and parameter values:

C	Accuracy
0.1	85.705%
6.08	85.007%
12.06	86.407%
12.11	85.68%
12.16	85.805%
12.20	85.233%
12.25	85.578%
12.26	84.853%
12.27	84.447%
12.28	85.943%
12.29	84.015%
12.30	86.443%
12.54	85.223%
12.78	85.128%
13.02	85.633%
13.26	85.458%
14.45	83.615%
15.65	85.378%
16.84	85.833%
18.04	86.262%
24.02	85.655%
30	84.123%

Average accuracy on test set with optimized parameter values:

Parameters:

10-fold CV

C = 12.30

Accuracy: **86.443%**