Series 2a

David Bücher, Timo Bürk, Félicien Hêche, Aleksandar Lazic, Zakhar Tymchenko

SVM

We choose the build our support vector machine (SVM) using Python and the scikit-learn library. We build this part of the project as follow :

- In the file SVM_linearVSgaussian.py we compare the performance of a support vector machine using a linear kernel and a support vector machine using a Gaussian kernel.
- In the file SVM_hyperparameter_tuning.py we try to optimize the hyperparameters of the model which gave the best result in the precedent file.
- Finally in the file SVM_test.py we compute the accuracy of our model on the test set.

Now, let's focus on the SVM_linearVSgaussian.py file. In this file, we want to compare two different SVM, one with a linear kernel and the other with the Gaussian kernel. To avoid too much computation time, we work only on a subset of the training set. (We use about the third of this set). And to measure the differences between our two models, we use a 5-fold cross validation procedure. We get the following output :

						Average
Linear kernel	0.91575	0.9185	0.91175	0.899	0.91425	0.91185
Gaussian kernel	0.95125	0.9455	0.9475	0.941	0.95175	0.94739999

We can observe that the model with Gaussian kernel provide a better result in all case. So, in the next file, we are going to tune this model to improve its accuracy.

Now, in the SVM_hyperparameter_tuning.py file, we try to find the optimal parameter of a support vector machine model using a Gaussian kernel. To do it, we test a C value of $1, 10, 100, 1000$, for the γ value, we try the value $0.001, 0.01, 0.1, 1$ and the standard γ value which is $1/n\_features = 1/785$ in our case. We make, like in the previous file and 5-fold cross-validation test with these different parameters and we get the following average :

	$\gamma =$ auto	$\gamma = 0.001$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
C = 1	0.947449999	0.945750000	0.791550	0.17665	0.1140500000
C = 10	0.9560500	0.9561	0.805999	0.1806	0.114050000
C = 100	0.9554	0.95475	0.80599999	0.1806	0.1140500
C = 1000	0.9554	0.95475	0.8059999	0.1806	0.11405000

We can observe that $\gamma = 0.01$ and $C = 10$ provide the best performance which is 0.9561

Finally in the SVM_test.py file, we build a support vector machine model using a Gaussian kernel, a C value of 10, a γ value of 0.001 and the all training set. After that we test it on the test set and we get an accuracy of 0.9724.