Machine Learning I – Exercise Sheet 2

Jan Zwank

1 EMG Physical Activation Dataset

We begin by loading the data from `http://archive.ics.uci.edu/ml/datasets/EMG+Physical+Action+Data+Set` with the corresponding labels and splitting it into a training set (80%) and a test set (20%).

We use this data to train both a support vector classifier and a random forest classifier. For that we use the `sklearn.svm.SVC` and `sklearn.ensemble.RandomForestClassifier` respectively from scikit-learn [1]. Here the random forests are applied for 10, 50, 100 and 500 trees.

Let's first consider the support vector classifier. We use the radial basis function as a kernel, since the data does not appear to be linearly separable. We also increase the cache size to help speed up the learning. If we consider the entire dataset, the learning takes 81.8 minutes. To evaluate the model, we consider the accuracy, sensitivity and specificity, as well as the receiver operating characteristic (ROC) curve and the area under the curve (AUC). These numerical metrics are included in Table 1. The ROC is provided in Figure 1.

For the random forest classifier, we do the same. The numerical metrics are included in Table 2 and the ROC is provided in Figure 2-5.

We can see that the random forest classifier is not only a lot faster, but also more accurate – even if only a small number of trees are used. Also note that the accuracy of our predictions hardly improve when more than 50 trees are used (For the out-of-bag-error refer to Figure 6). However, the accuracy does not decrease either. The training time, however increases with the number of trees.

accuracy	0.8110
sensitivity	0.80
specificity	0.82
AUC	0.8110

Table 1: Numerical metrics for SVC

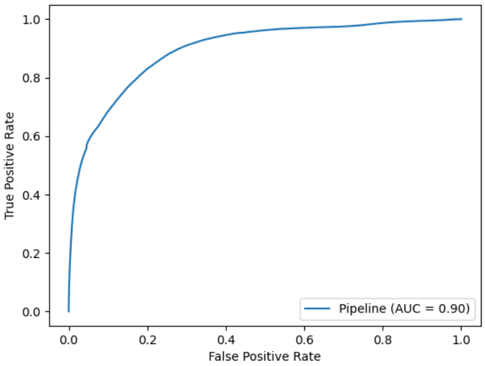

Fig. 1: ROC curve for SVC

metric	10 Trees	50 Trees	100 Trees	500 Trees
accuracy	0.9646	0.9716	0.9714	0.9723
sensitivity	0.97	0.98	0.98	0.98
specificity	0.96	0.96	0.96	0.96
AUC	0.9646	0.9716	0.9713	0.9723

Table 2: numerical metrics for random forests

2 Breast Tissue Dataset

We begin by loading the dataset from `http://archive.ics.uci.edu/ml/datasets/Breast+Tissue` and merge the classes ”fad”, ”mas” and ”gla” into one single class.

We first apply a ”one-vs-rest” classifier in connection with a support vector classifier using a linear, sigmoid and Gaussian kernel.

linear: The resulting ROC with some extensions can be found in Figure 7. The accuracy for the linear kernel was 77.3%. Further metrics are given in Table 3. Note that specificity and sensitivity were replaces with precision and recall of the corresponding classes.

sigmoid: The resulting ROC with some extensions can be found in Figure 8. The accuracy for the linear kernel was 69.8%. Further metrics are given in Table 4.

metric	fad/mas/gla	adi	car	con
precision	1.00	0.91	0.69	1.00
recall	0.84	1.00	0.90	0.50
AUC	0.99	1.00	0.98	0.98

Table 3: Numerical metrics for SVC with linear kernel

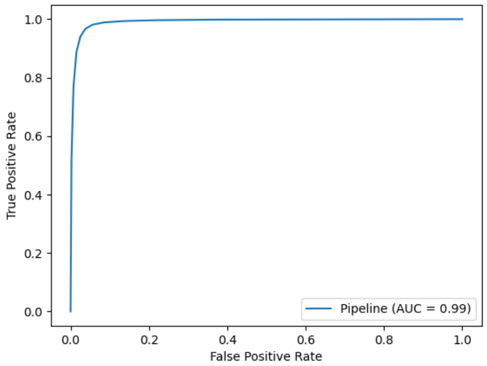

Fig. 2: ROC curve for random forest with 10 trees

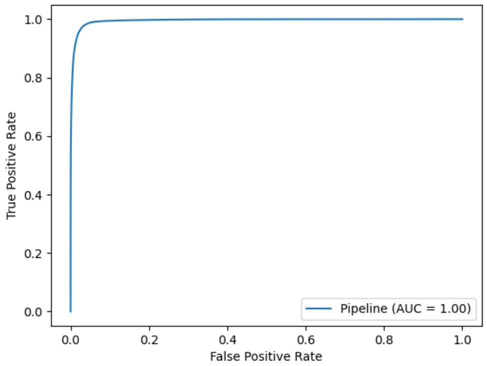

Fig. 3: ROC curve for random forest with 50 trees

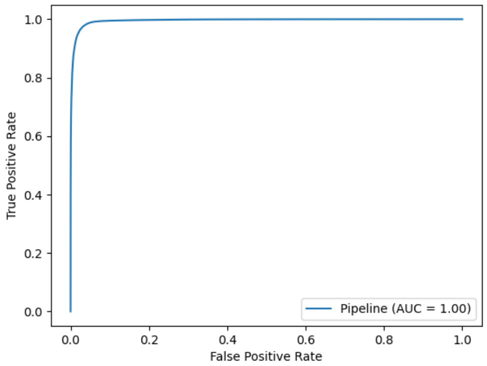

Fig. 4: ROC curve for random forest with 100 trees

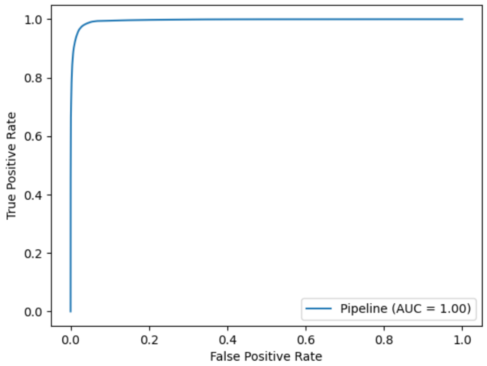

Fig. 5: ROC curve for random forest with 500 trees

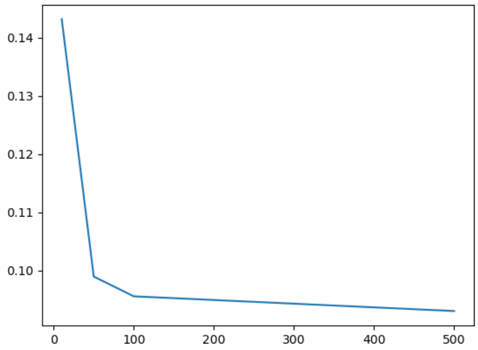

Fig. 6: Out-of-bag-error for Random Forest with 10, 50, 100 and 500 trees.

metric	fad/mas/gla	adi	car	con
precision	0.92	0.67	0.67	0.00
recall	0.96	1.00	0.60	0.00
AUC	0.99	1.00	0.93	0.61

Table 4: Numerical metrics for SVC with sigmoid kernel

metric	fad/mas/gla	adi	car	con
precision	0.92	0.91	0.82	1.00
recall	0.96	1.00	0.90	0.12
AUC	0.98	1.00	0.99	1.00

Table 5: Numerical metrics for SVC with Gaussian kernel

Note that specificity and sensitivity were replaces with precision and recall of the corresponding classes.

Gaussian: The resulting ROC with some extensions can be found in Figure 9. The accuracy for the linear kernel was 81.1%. Further metrics are given in Table 5. Note that specificity and sensitivity were replaces with precision and recall of the corresponding classes.

It appears, that class 'con' is particularly difficult to classify correctly when using support vector classifiers. From all SVCs the Guassian kernel performs the best and the sigmoid kernel the worst.

Now using a random forest classifier with 100 trees, we get the ROC in Figure 10. The `classification_report` indicates, that the random forest classifies all testing samples correctly, i.e. the accuracy is 100%. This also means that all precision and recall metrics are equal to 1.

3 p53 Mutants Dataset

We start by loading the K8 dataset of `p53_old_2010` from `http://archive.ics.uci.edu/ml/datasets/p53+Mutants` and classify them as *active* or *inactive* with both SVC and random forest. For the random forest we compare the results for 100, 500, 1000, 2000 and 5408 trees. The numerical metrics can be found in Table 6 and the ROCs can be found in Figures 11-16. The plot of the out-of-bag-errors for the random forests are provided in Figure 17. The increase in the out-of-box-error for $N > 1000$ indicates that the model is overfitting the data and less trees should be used to get optimal results.

References

[1] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

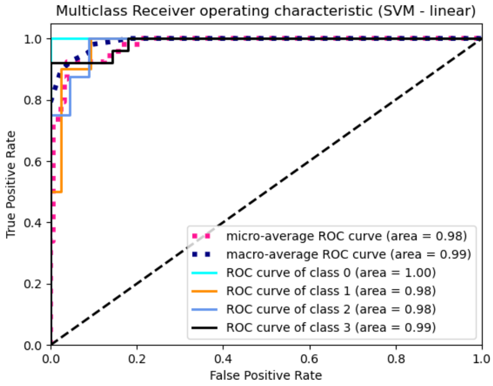

Fig. 7: ROC for SVC with linear kernel,
0:'adi',
1:'car',
2:'con',
3:'fad/mas/gla'

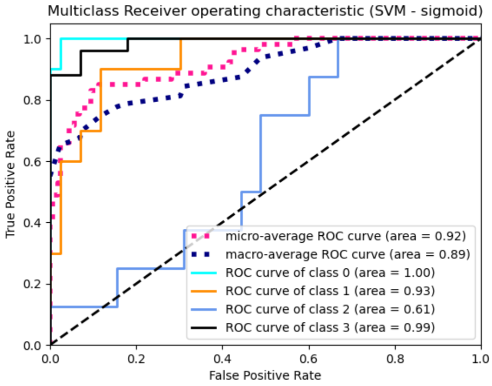

Fig. 8: ROC for SVC with sigmoid kernel,
0:'adi',
1:'car',
2:'con',
3:'fad/mas/gla'

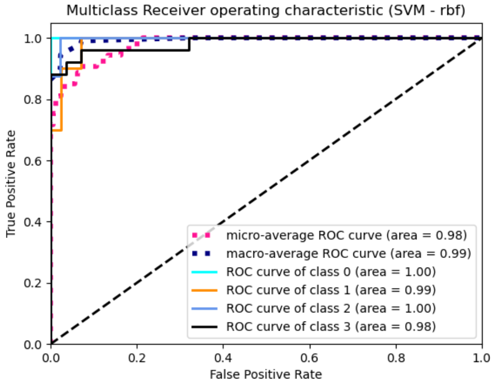

Fig. 9: ROC for SVC with rbf kernel,
0:'adi',
1:'car',
2:'con',
3:'fad/mas/gla'

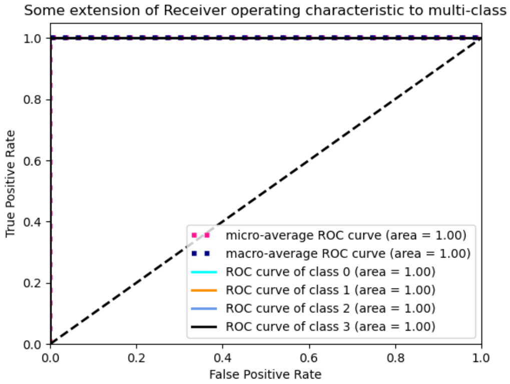

Fig. 10: ROC for RF,
0:'adi',
1:'car',
2:'con',
3:'fad/mas/gla'

Model	SVC	RF100	RF500	RF1000	RF2000	RF5408
Accuracy	0.9964	0.9982	0.9984	0.9984	0.9984	0.9984
Sensitivity	0.82	0.82	0.82	0.82	0.82	0.80
Specificity	1.00	1.00	1.00	1.00	1.00	1.00
AUC	0.97	0.98	0.99	1.00	1.00	1.00

Table 6: Numerical metrics for various classifications.

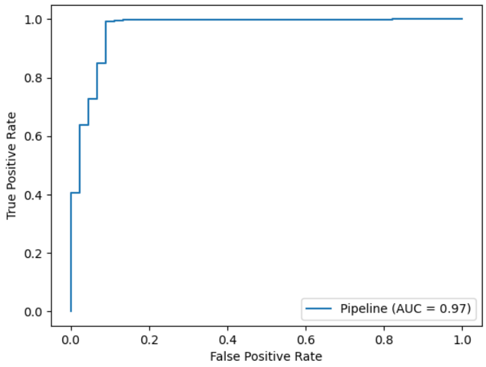

Fig. 11: ROC for SVC

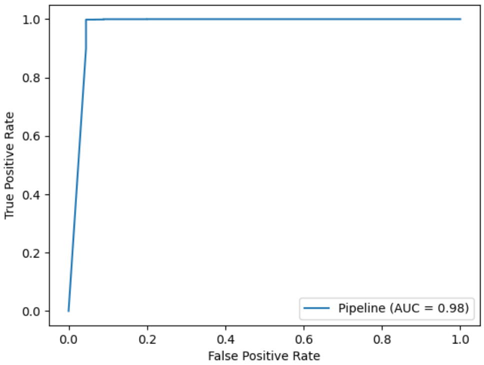

Fig. 12: ROC for random forest with 100 trees.

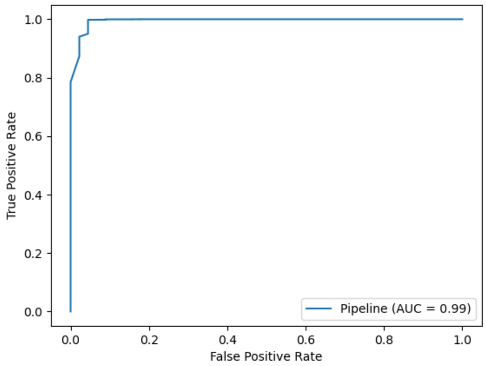

Fig. 13: ROC for random forest with 500 trees.

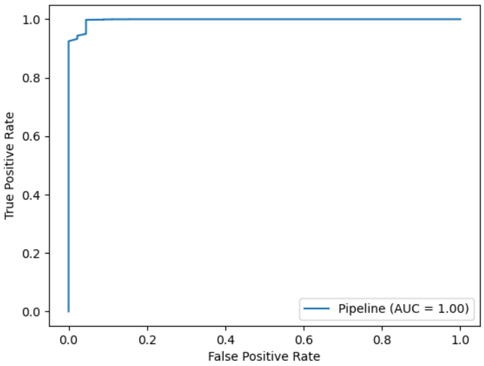

Fig. 14: ROC for random forest with 1000 trees.

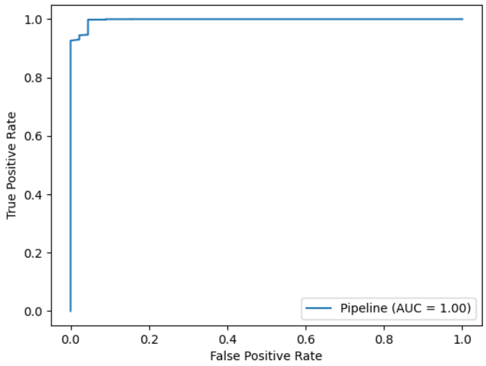

Fig. 15: ROC for random forest with 2000 trees.

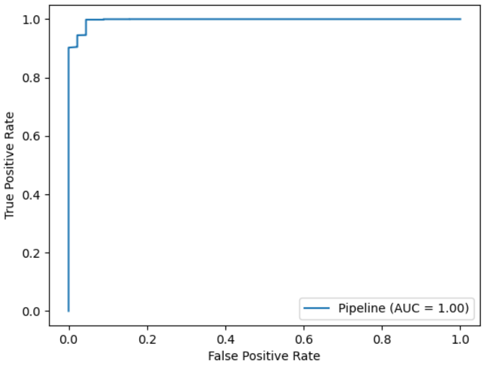

Fig. 16: ROC for random forest with 5408 trees.

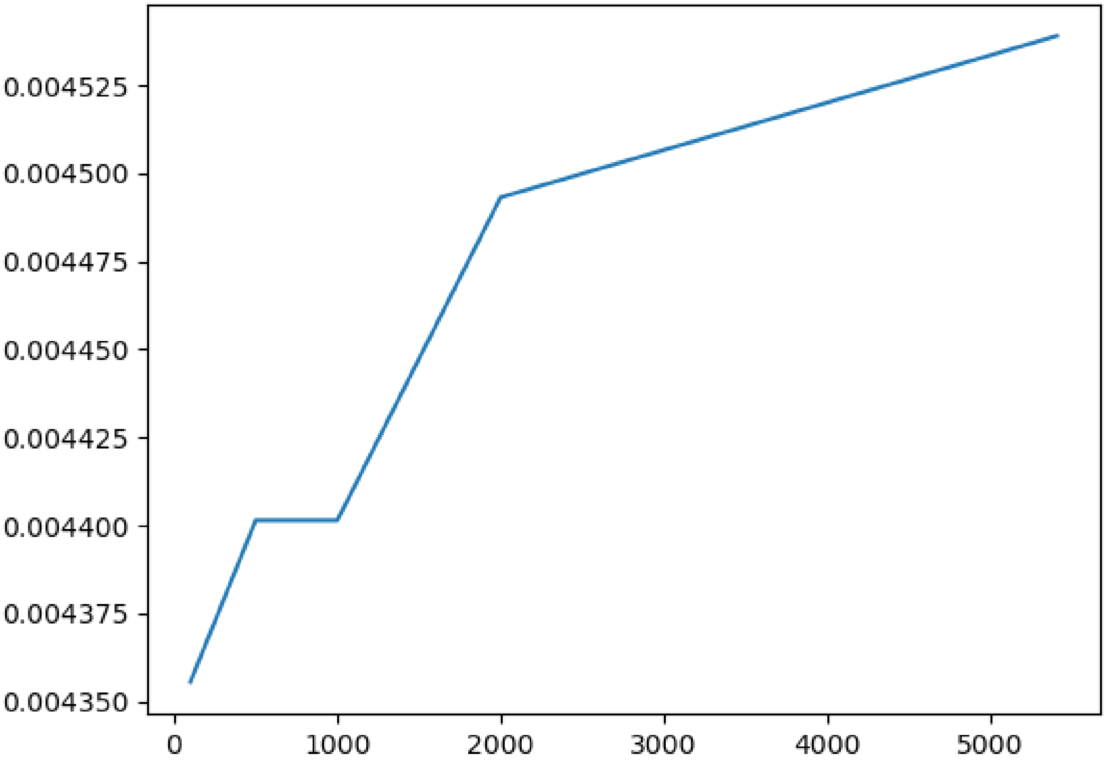

Fig. 17: Out-of-bag-error for random forests with 100, 500, 1000, 2000, 5408 Trees