Assignment 3

Computational Intelligence

Team Members		
Last name	First name	Matriculation Number
Merdes	Malte	01331649
Riedel	Stefan	01330219

1 Linear SVM

a)

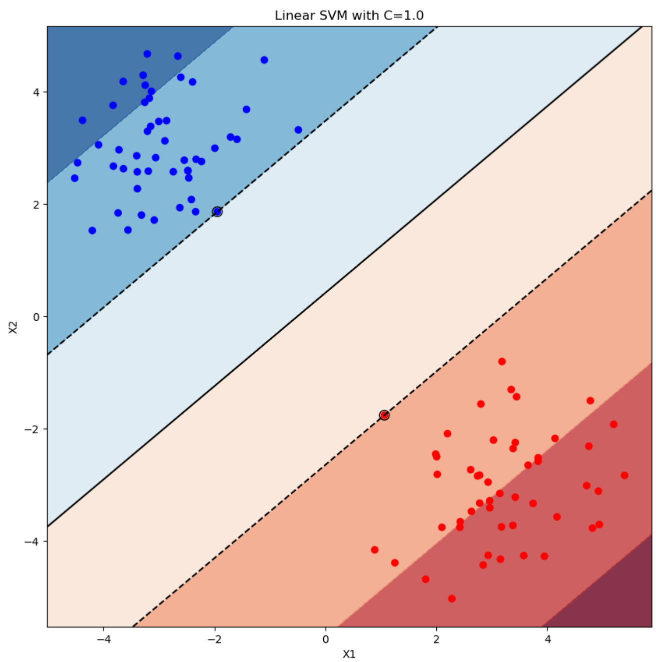

Figure 1: Linear SVM

b)

After adding the new outlier point (4,0) to the data set, we observe a rotated decision hyperplane and a smaller margin. Since the outlier (with label y = 1) is close to the points of the other class (y = -1) it leads to a worsened general prediction behaviour (figure 2).

c) The larger the parameter C, the more we penalize misclassification on the training and tend to overfit the training data (figures 3 and 4).
With a smaller value of C = 0.1 we get a larger error on the training, but avoid overfitting and achieve a better generalized prediction (figure 5).
However, if C gets too small, we underfit the data and missclassify completely (figure 7).

In general, smaller values for C lead to an increased number of support vectors as we allow for more points to be inside the margin.

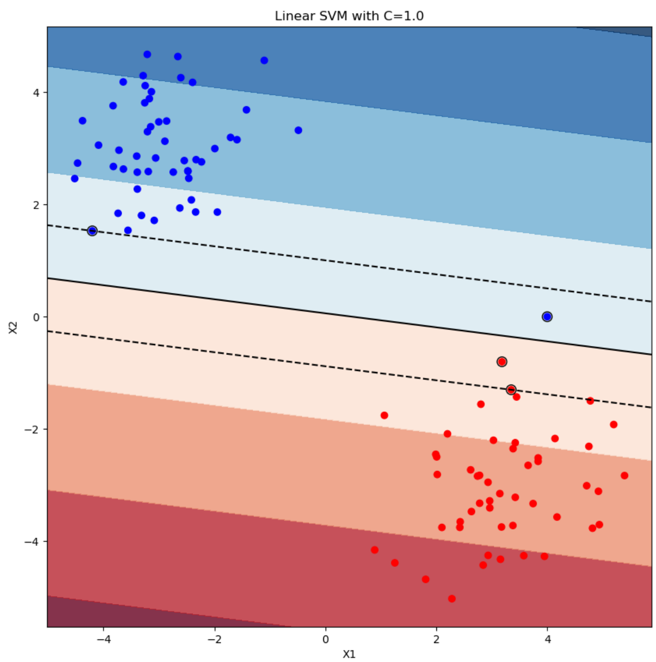

Figure 2: Linear SVM with added outlier point (4,0)

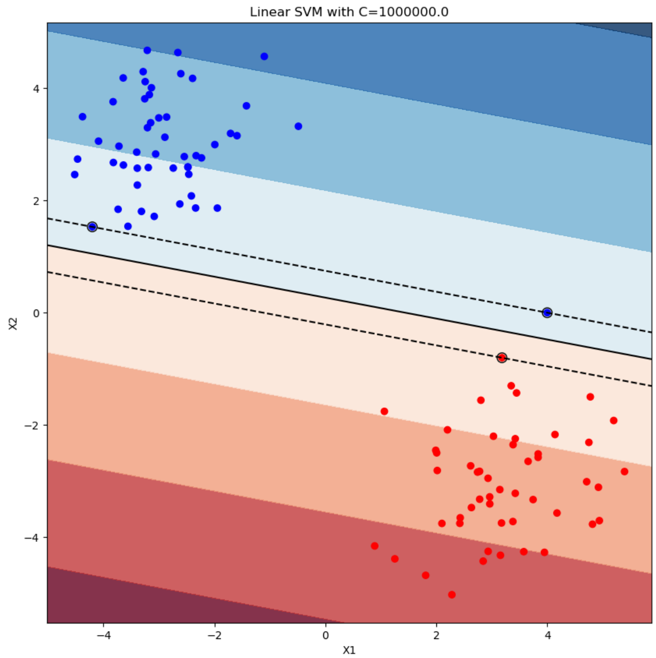

Figure 3: Linear SVM with C = 1000000

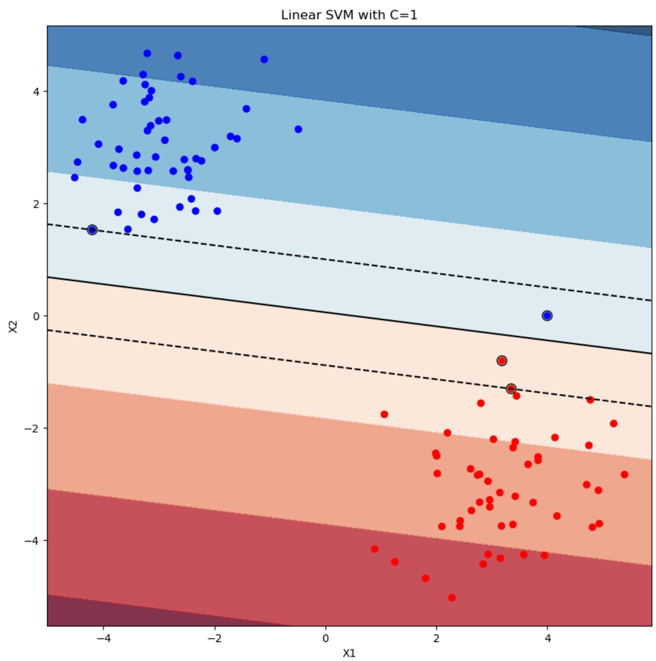

Figure 4: Linear SVM with C = 1

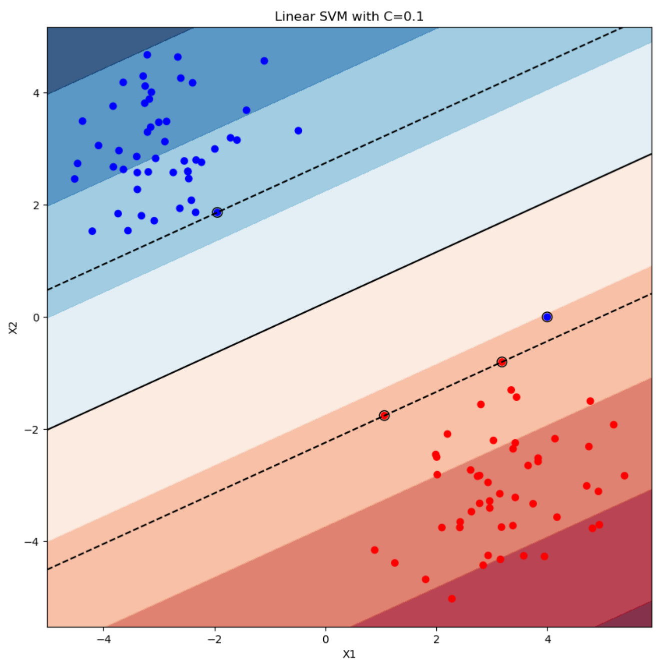

Figure 5: Linear SVM with C = 0.1

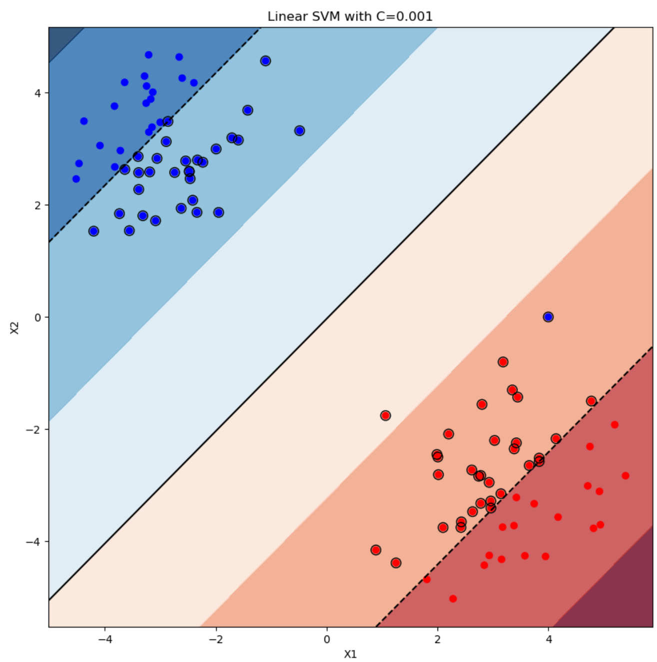

Figure 6: Linear SVM with C = 0.001

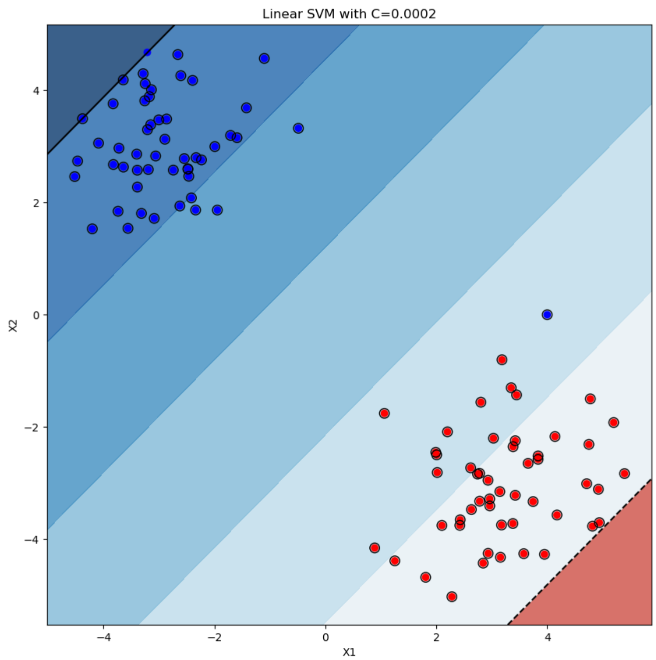

Figure 7: Linear SVM with C = 0.0002

2 Nonlinear (kernel) SVM

a)
First we train a linear kernel SVM on the nonlinear data and plot the decision boundary (fig. 8). Obviously, the linear kernel can't fit the data very well. This is confirmed with a classification score on the test set of only 0.8125.

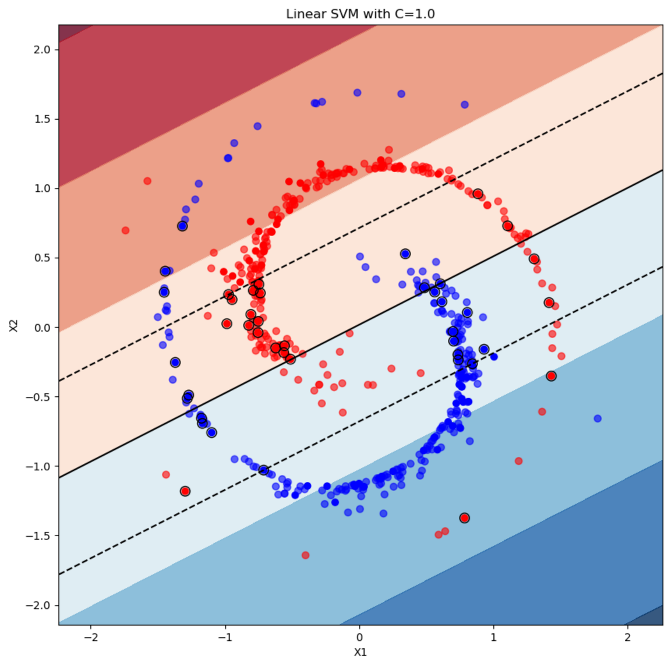

Figure 8: Linear SVM with C = 1

b)
Now we train a polynomial kernel SVM for degrees in range 1 to 20 and plot the classification score vs. degree (fig. 9). The best score was achieved for a polynomial degree of 9 with a classification score of 0.9575.

c)
Finally, we train a RBF kernel SVM with γ in range 0.01 to 1.99. For γ in range 1.85 to 1.99 we achieve the best test classification scores of 0.9425 (fig. 11).

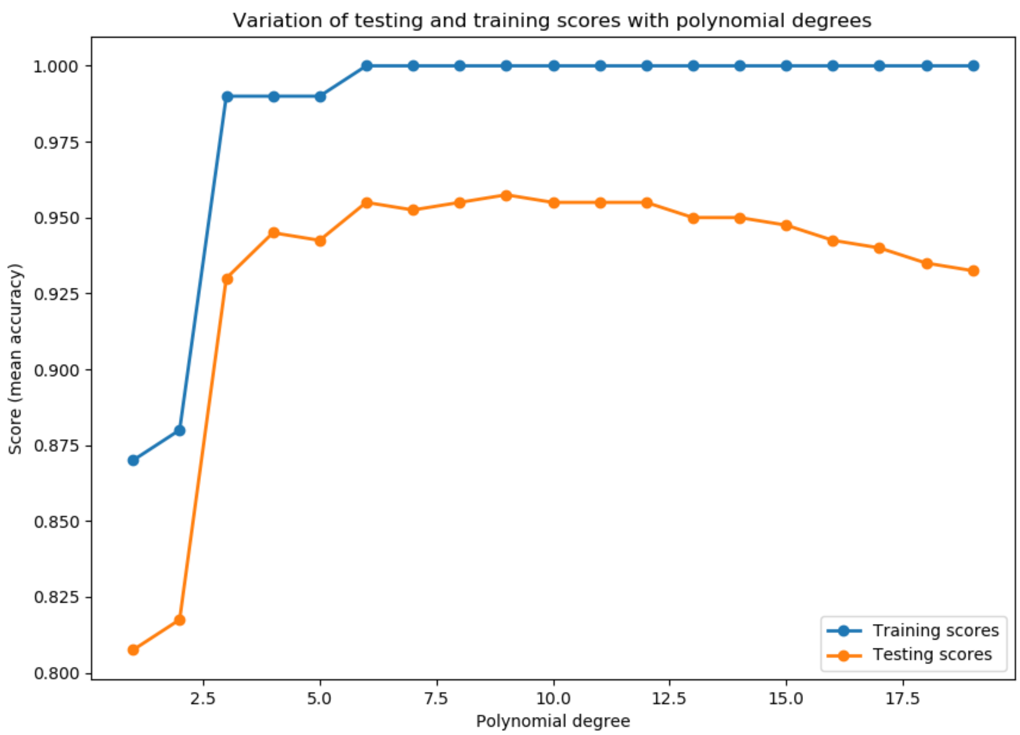

Figure 9: Classification score vs. polynomial kernel degree

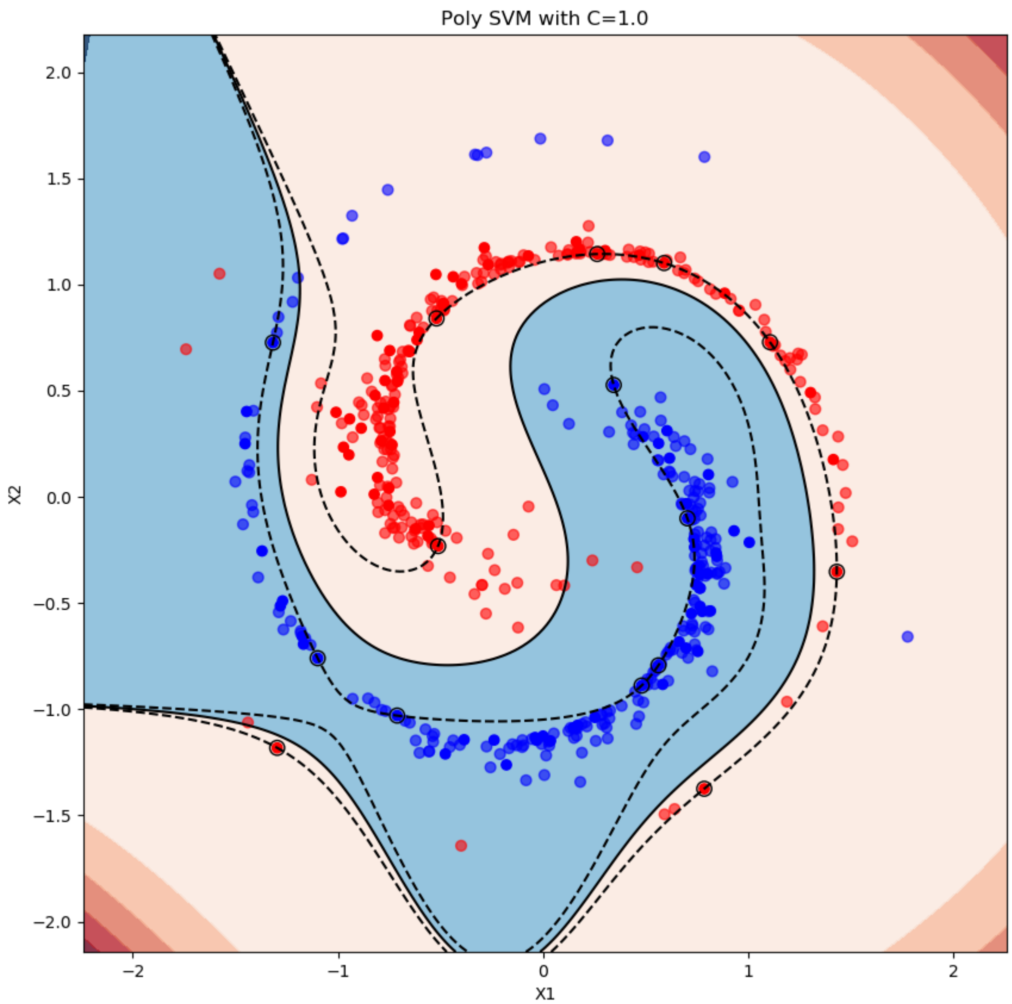

Figure 10: Decision boundary of optimal polynomial kernel with degree 9

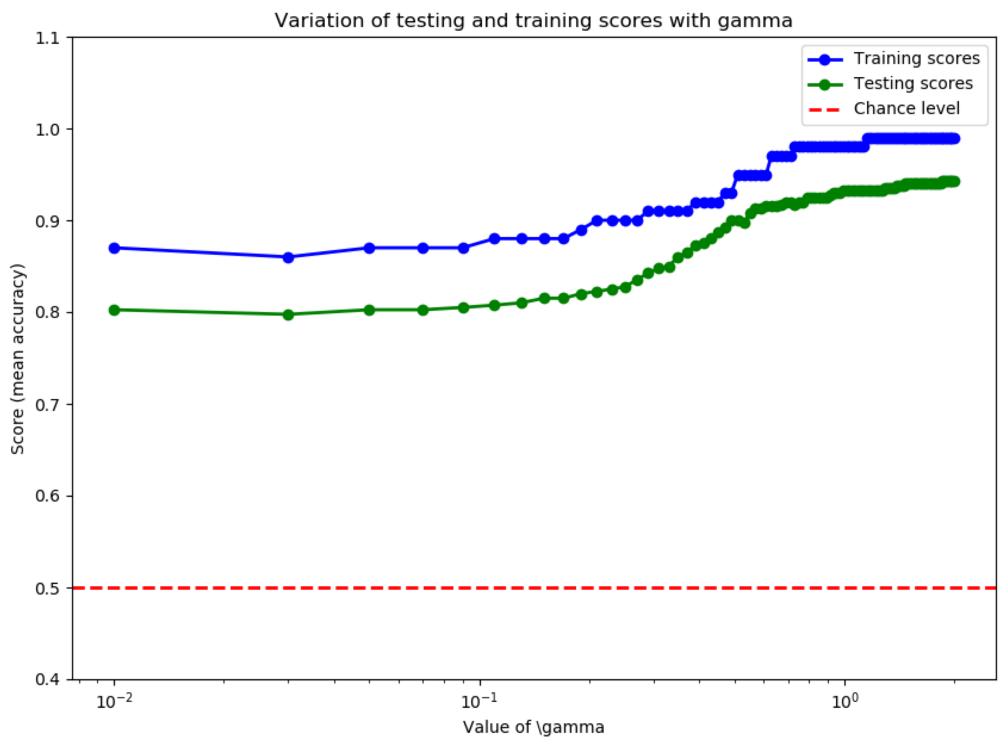

Figure 11: Classification score vs. parameter γ of RBF kernel

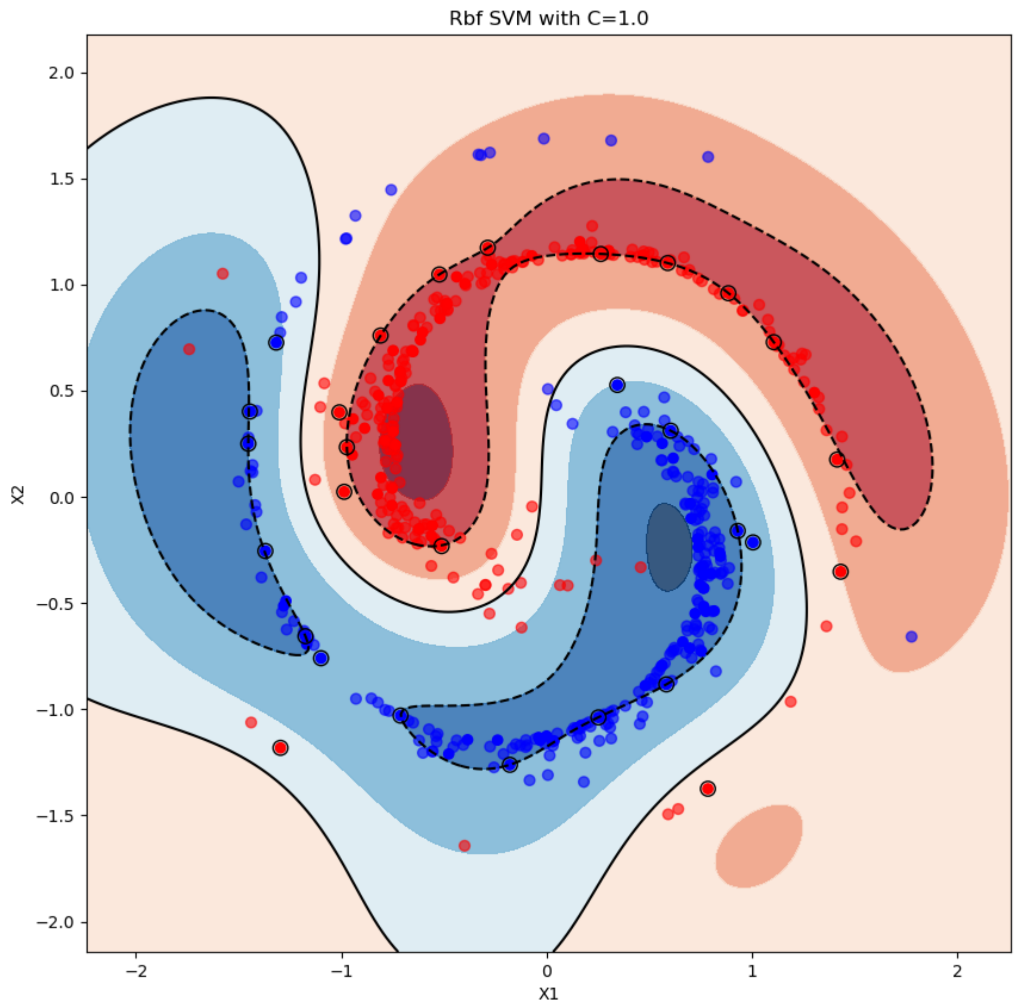

Figure 12: Decision boundary of optimal RBF kernel with $\gamma = 1.85$

Comparison:

The best performing kernel was the polynomial kernel of degree 9, which achieved a test score of 0.9575. In contrast, the RBF kernel SVM achieves the better generalization already for low values of gamma, whereas the polynomial SVM requires a high degree to achieve reasonable scores. The complexity of the decision boundaries is comparable for both the polynomial and RBF kernels, although the number of support vectors is higher for the RBF kernels.

3 Multiclass classification

a)
'One-versus-Rest' algorithms split the problem into several binary classifications. In our case, a 5-class problem is reduced into 5 binary classifications:
1 vs. (2345) , 2 vs. (1345) and so on.
'One-versus-one' algorithms split the multiclass problem into several real binary classifications. In our case, a 5-class problem is reduced into $\frac{5*(5-1)}{2} = 10$ binary classifications:
1 vs. 2 , 1 vs. 3, ... , 2 vs. 1 and so on.
The linear kernel performs well on text images, because text is inherently discrete (either white or black pixels), sparse, and high dimensional (many pixels) data. This makes it easy to fit a hyperplane reasonably well though the data space for binary classification.

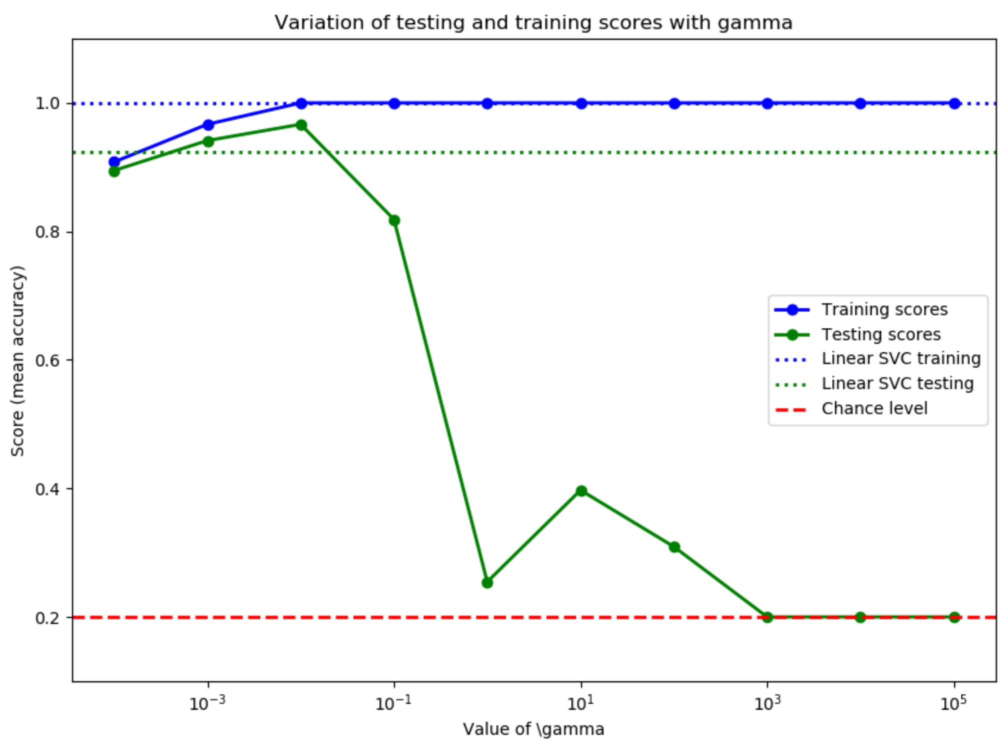

Figure 13: Comparison of RBF kernel (as function of gamma) versus a linear kernel. The RBF kernel outperforms the linear for gamma = 10^{-2}.

b) According to the following confusion matrix, we found the highest error rate for class 5 (error-rate = 0.12258065, see python for the calculation).

Confusion matrix:
[147 1 2 0 0]
[3 142 0 4 1]
[4 2 126 0 18]
[3 5 0 142 0]
[2 3 9 0 136]

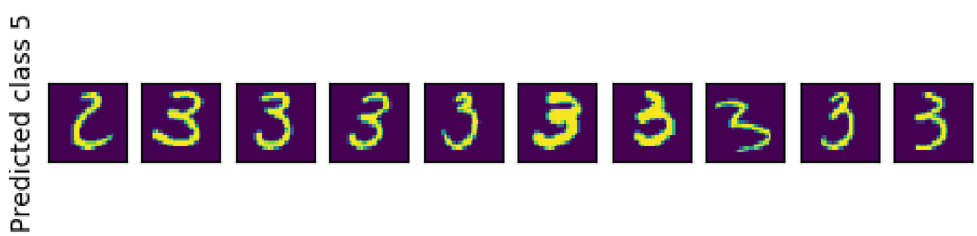

Figure 14: Misclassified pictures (classified as class 5)

Looking at the the missclassified pictures we observe that the SVM often misclassifies 3 as 5, because the lower half of the handwritten digit is very similar. This is also confirmed by the confusion matrix, which has the highest misclassification entry (18) for the predicted class 5, when it was actually 3.