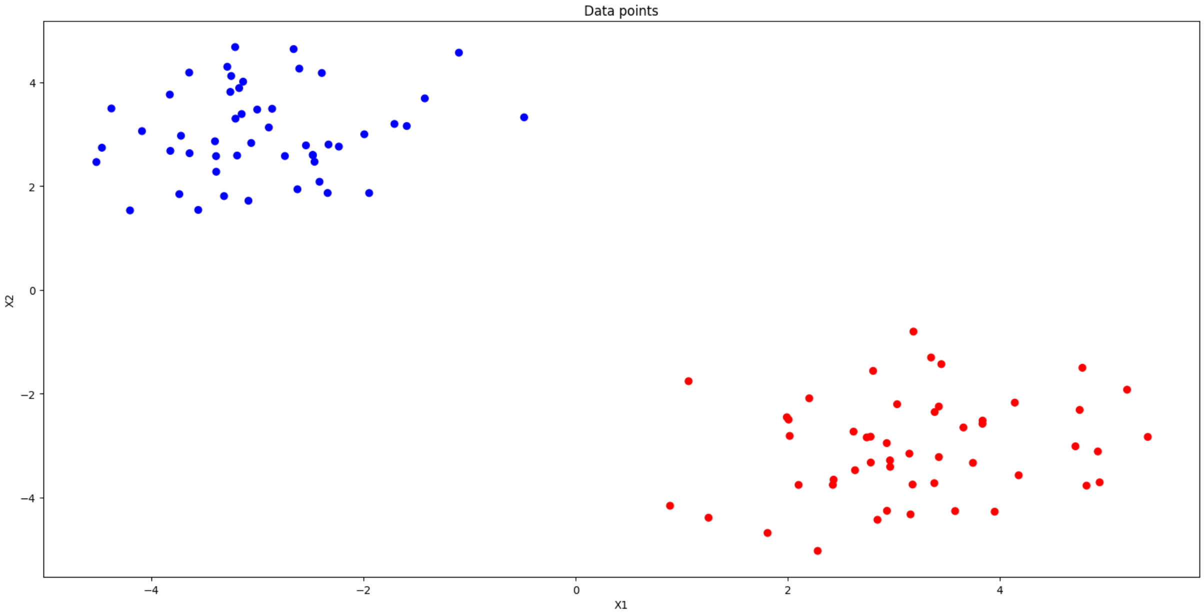
Assignment 3

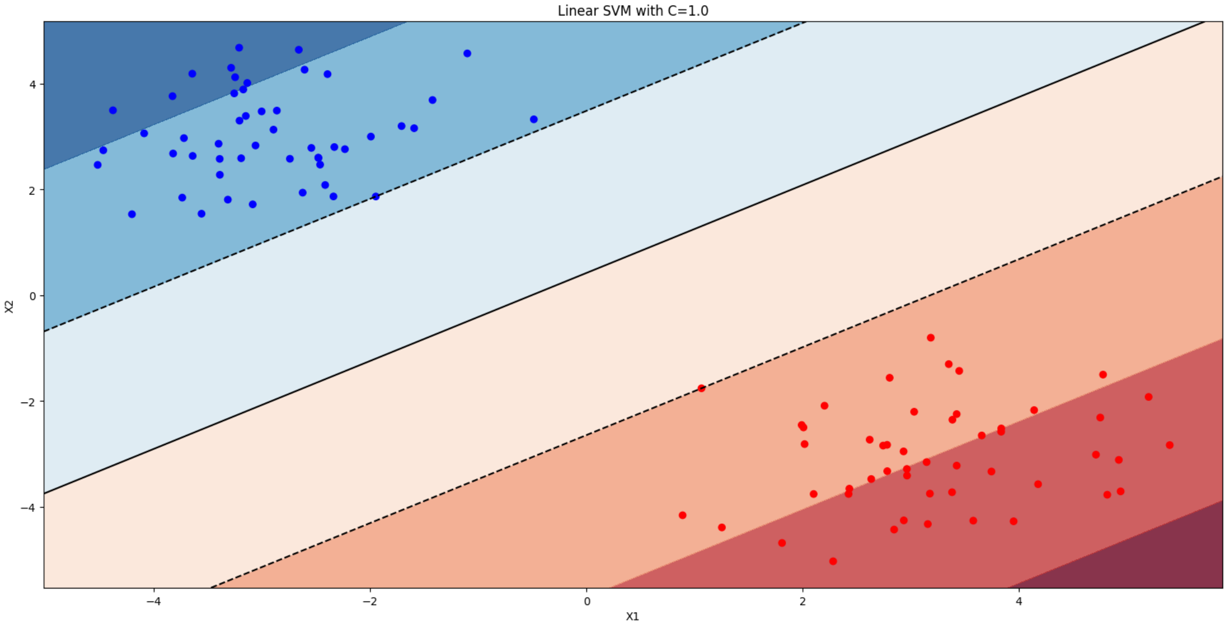
Computational Intelligence SEW, SS2017

Team Members		
Last name	First name	Matriculation Number
Papst	Stefan	1430868
Guggi	Simon	1430534
Perkonigg	Michelle	1430153

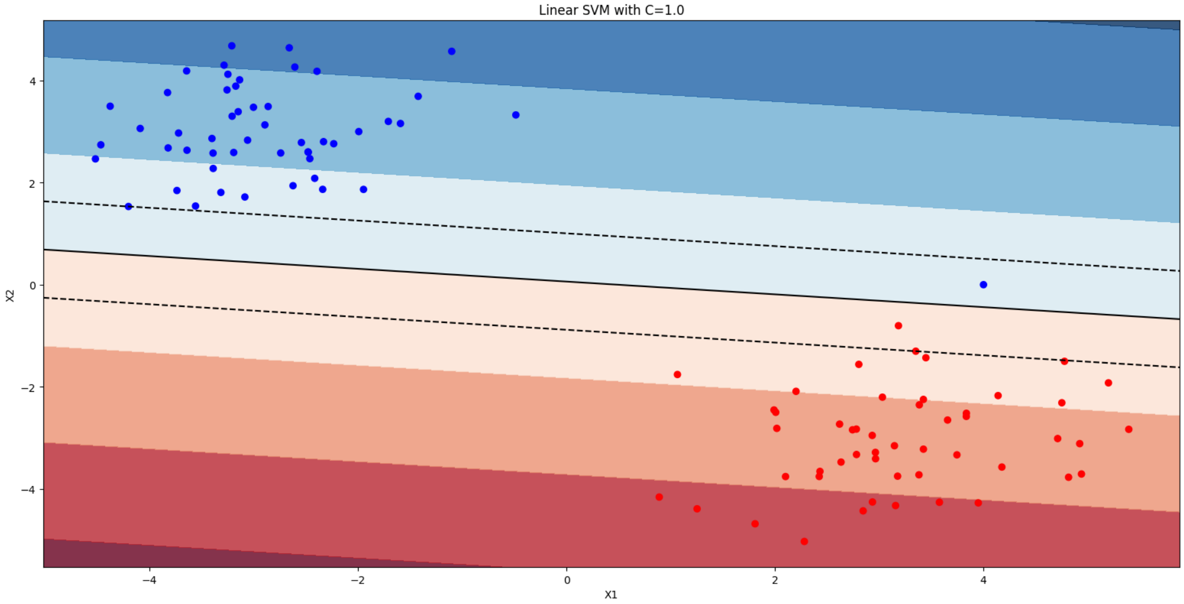
1 Linear SVM

Include plots of all the results

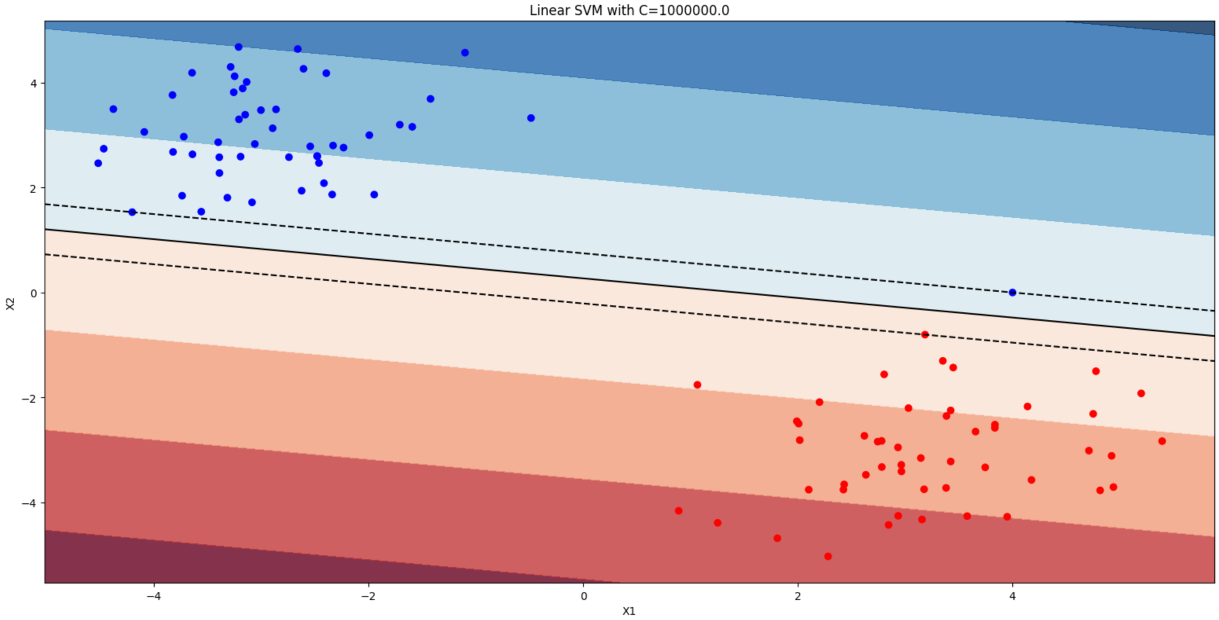


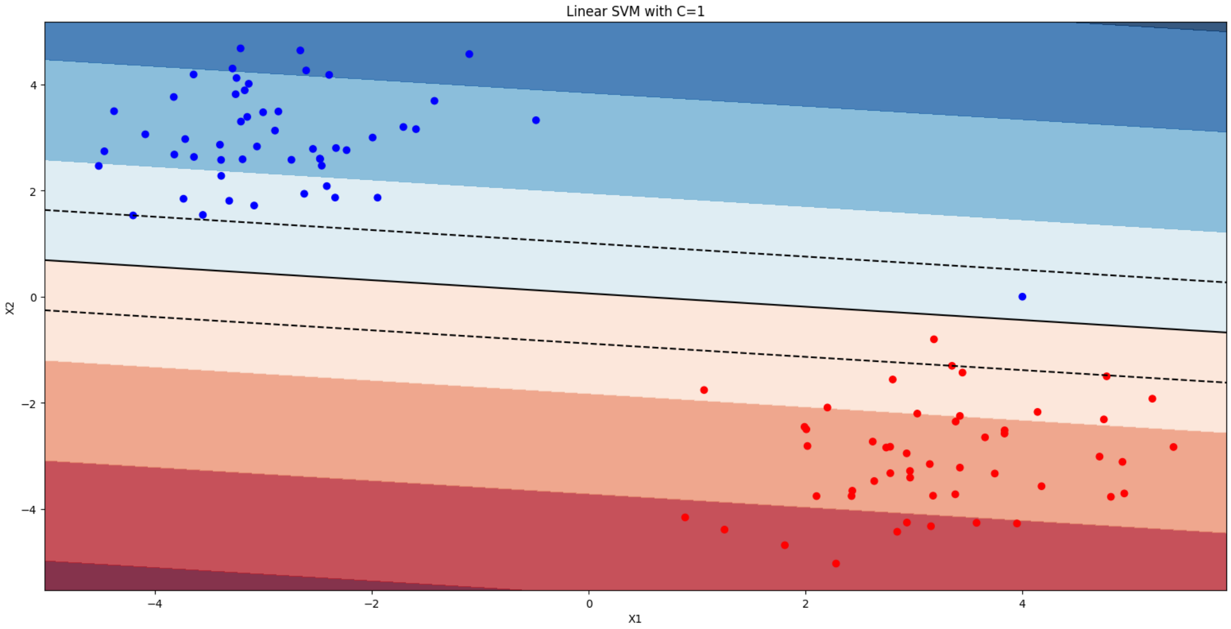


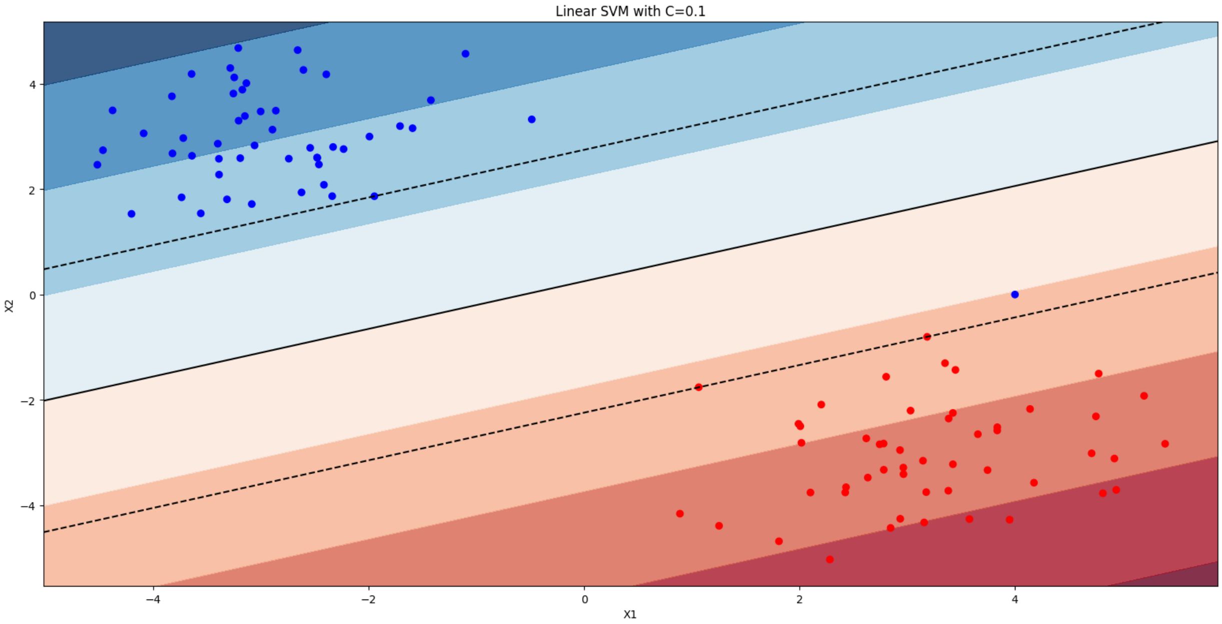
Task b:

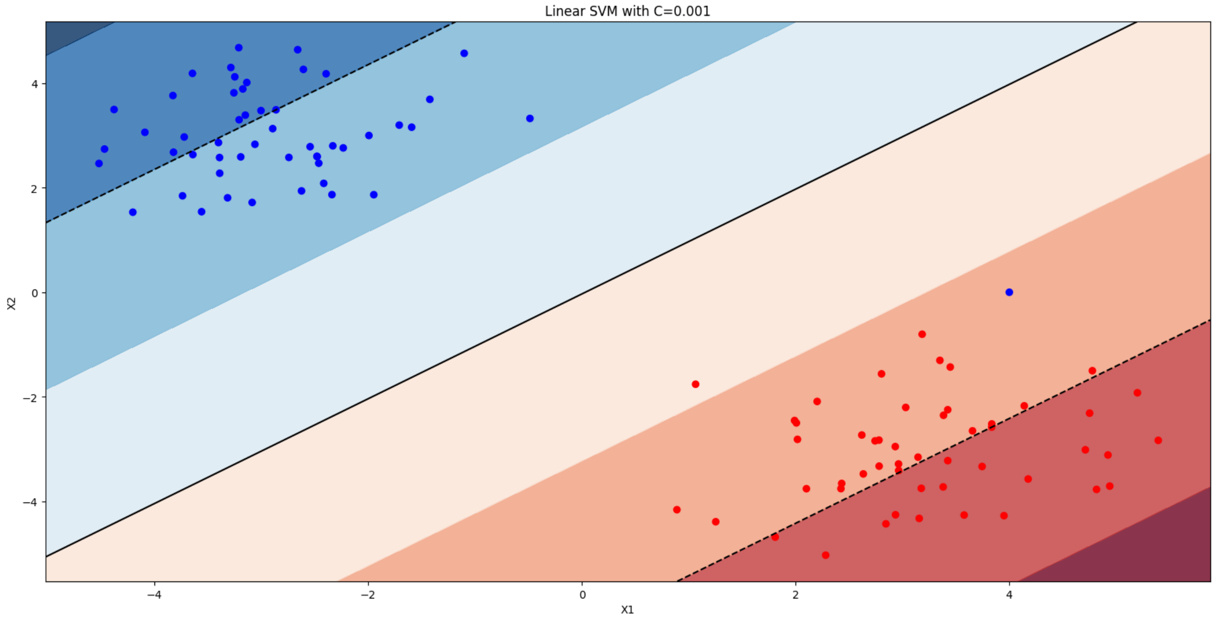

Linear SVM with C=1.0
X2
4
2
0
−2
−4
X1
−4
−2
0
2
4

Task c:


Linear SVM with C=1000000.0
X2
4
2
0
−2
−4
X1
−4
−2
0
2
4


Linear SVM with C=1
X2
X1
4
2
0
−2
−4
−4
−2
0
2
4


Linear SVM with C=0.1
X2
X1
4
2
0
−2
−4
−4
−2
0
2
4



For task b), discuss how and why the decision boundary changed when the new point was added.

The SVM tries to find a decision boundary that separates the classes. When adding the new point, the blue point, on the specific position (4, 0) that is very isolated from the rest of the points of the blue class the decision boundary must be set new including this significant point. This can be seen at the third plot.

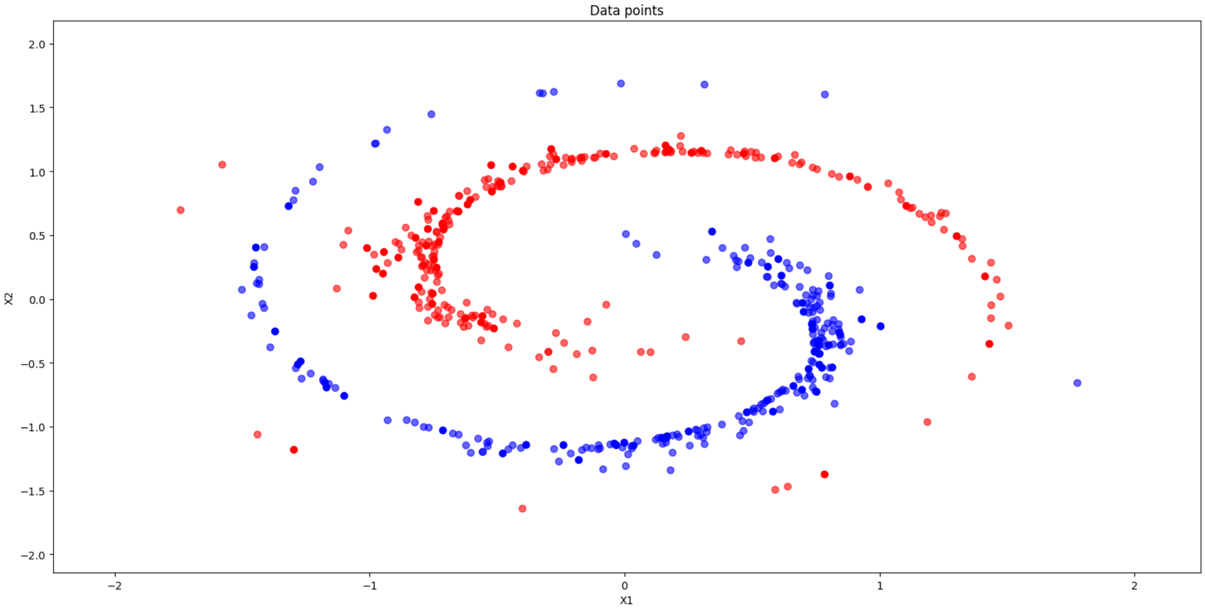
For task c):

- *Report how the parameter C influences the decision boundary found by the SVM*
  The parameter *C* determines whether you want to find a hyperplane with the largest minimum margin or a hyperplane that separates the samples correctly (as many as possible). Thus, we can say that the parameter *C* describes the sensitivity of the SVM to outliers like the new point added. A high *C* represents high sensitivity and (almost) correct classification, whereas a low *C* represents low sensitivity and the possibility for misclassification.
- *How does the number of support vectors found by the SVM change with the value of C? Why?*
  The intention of using a SVM is given by two reasons: a hyperplane with the largest minimum margin and a hyperplane, which separates the dataset as correctly as possible. The value of C stands for the desire, which of the two reasons should be prioritized. So a

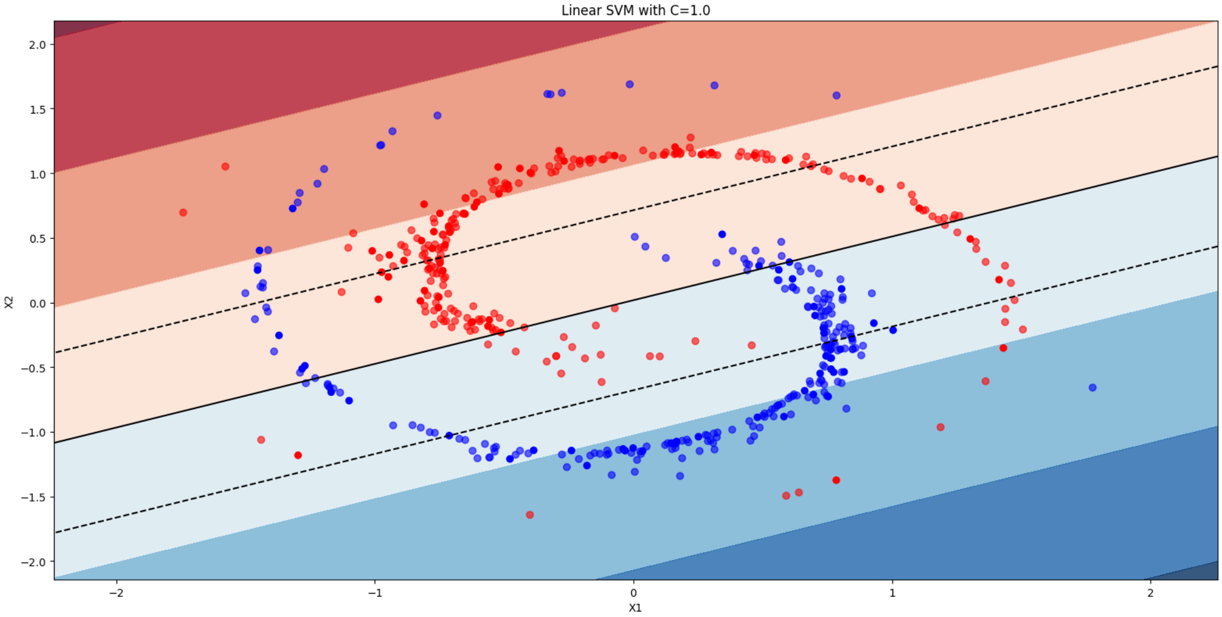
lower value of C gives one a very large minimum margin, which increases the number of the support vectors.

2 Nonlinear (kernel) SVM

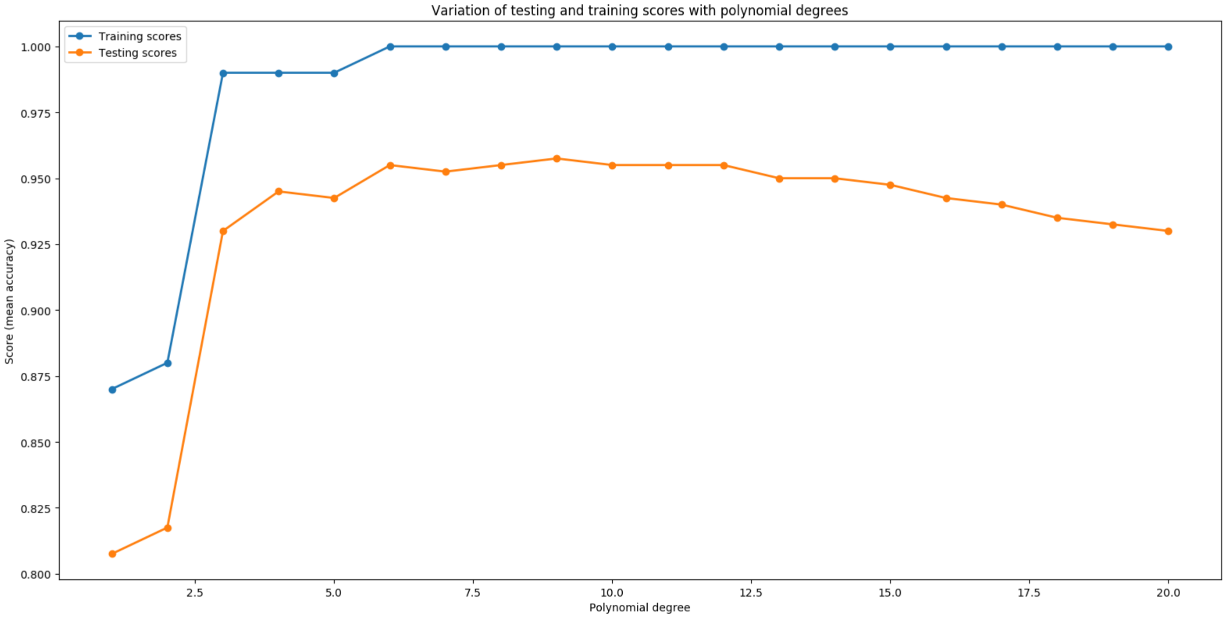
Include plots of all the results

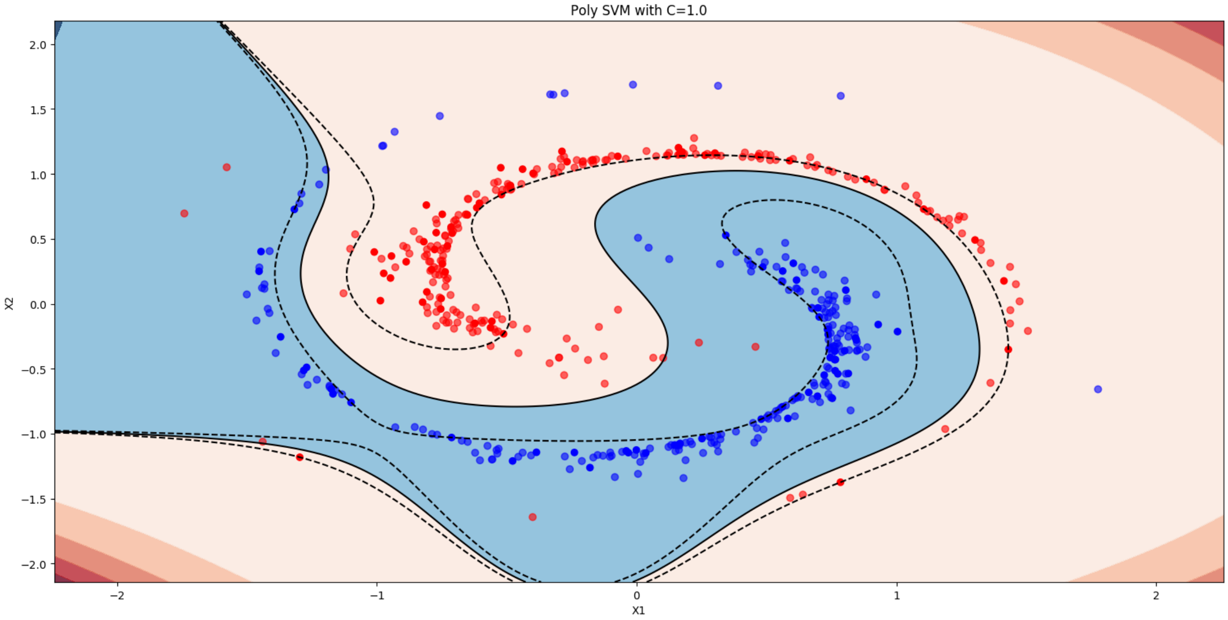


Task a:

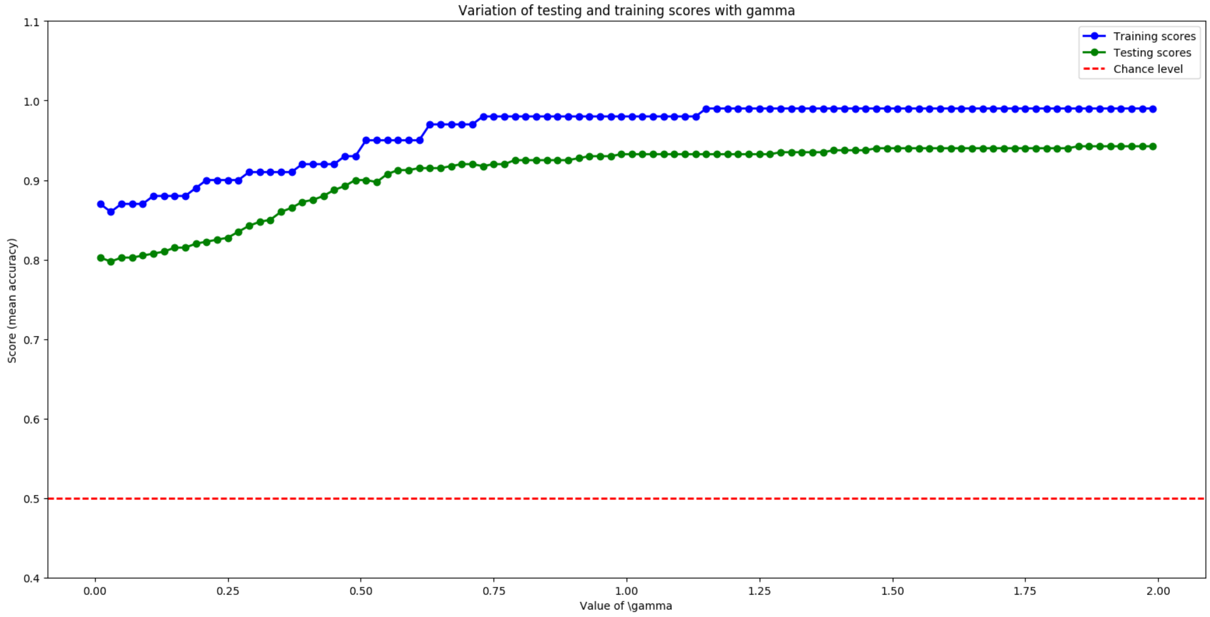


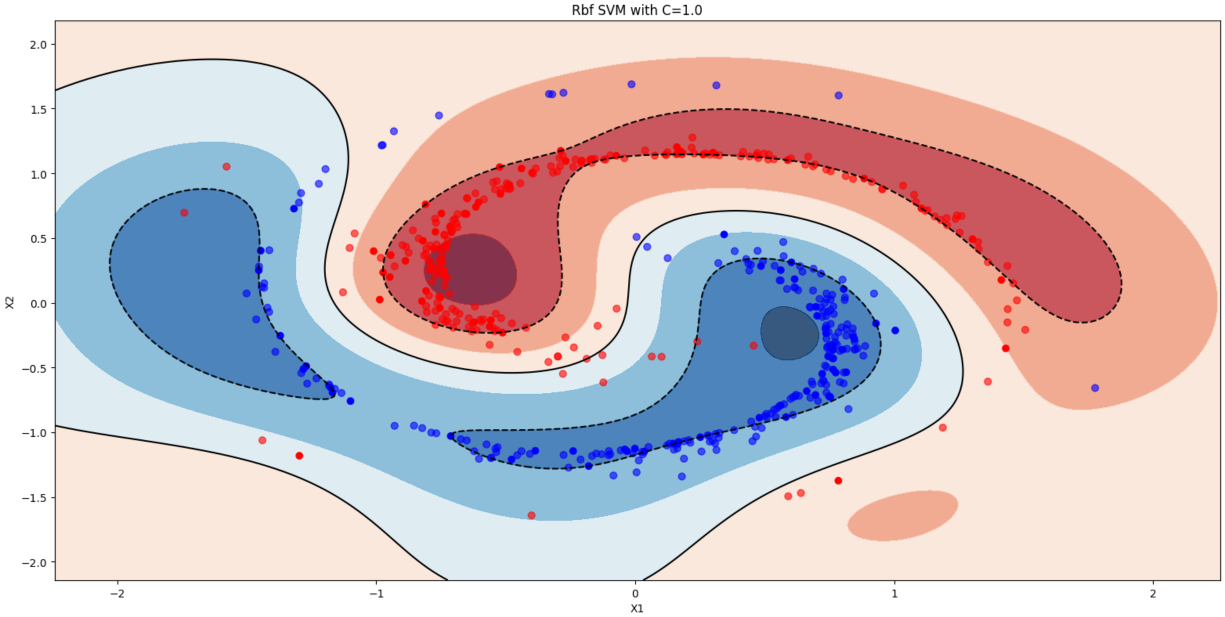
Task b:


Variation of testing and training scores with polynomial degrees
Training scores
Testing scores
Score (mean accuracy)
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Polynomial degree


Poly SVM with C=1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
X2
−2
−1
0
1
2
X1

Task c:





For task b), which degree of the polynomial produces the highest test score (accuracy)? Report this test score.

degree for highest test score: 9

highest test score for poly kernel: 0.9575

For task c), which value of gamma produces the highest test score (accuracy)? Report this test score.

gamma of highest test score: 1.85

highest test score for rbf kernel: 0.9425

Compare results obtained by each of these three kernels:

- *State the maximum test score achieved for each of these kernels and the kernel parameter for which that was achieved.*
  test score for **linear kernel** 0.8125
  highest test score for **poly kernel** 0.9575 with degree 9
  highest test score for **rbf kernel** 0.9425 with gamma 1.85
- *Which of the considered kernels performs best and why?*
  The polynomial kernel gives us the best results, but the linear kernel is faster.
- *Compare the complexity of decision boundaries and the number of Support Vectors found.*
- The linear SVM has a shape of (43,2), the polynomial SVM has (20,2) and the rbf one has (29,2). The linear SVM has the most rows so it is the most complex one.
- *Which kernel generalizes best for the given dataset?*
- In this special case we would suggest the polynomial kernel, because it has the best test score and the lowest number of rows in the support vector matrix. Because of this small dataset, the speed for the calculation is not as important as for other datasets. Almost all three kernel terminate in the same time.

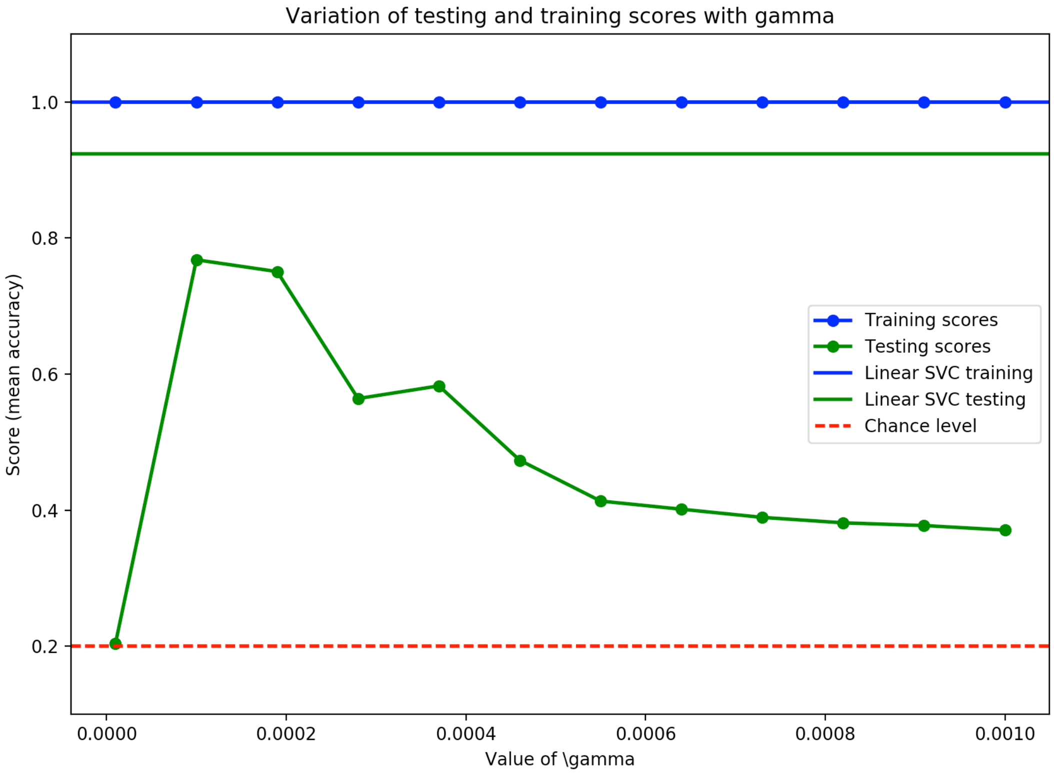
3 Multiclass classification

Recall the algorithms `One-versus-Rest' (or versus-all) and `One-versus-one' multi-class classification procedures. How many binary classifiers need to be trained in both cases?

OVA creates a classifier that distinguish each class from all others. For each class there is one classifier. At the end the prediction with the highest confidence score is selected.
OVO creates a classifier that distinguish between each pair of classes. So there are N(N-1)/2 classifier for N classes. After training, in the testing part the class with the most votes is chosen.
One-versus-Rest: N classifiers for N classes (one per class)
One-versus-One: N (N-1) / 2 classifiers for N classes

Include plots for ex_3_a with the scores of a linear and a rbf kernels.

Best linear test score : 0.924
Best rbf test score: 0.768 with gamma: 0.0001



Discuss those results. In particular why does a linear kernel perform well on images?

The straight green line shows the testing score of the linear SVC. The green line with points in it is obviously the score for the rbf SVC, because the peak at gamma 0.0001 agrees with the highest score from the trained rbfSVM. This result means to us that the dataset is pretty clean linear separable

As you can see, the linear kernel is much better than the rbf kernel. The plot shows the low

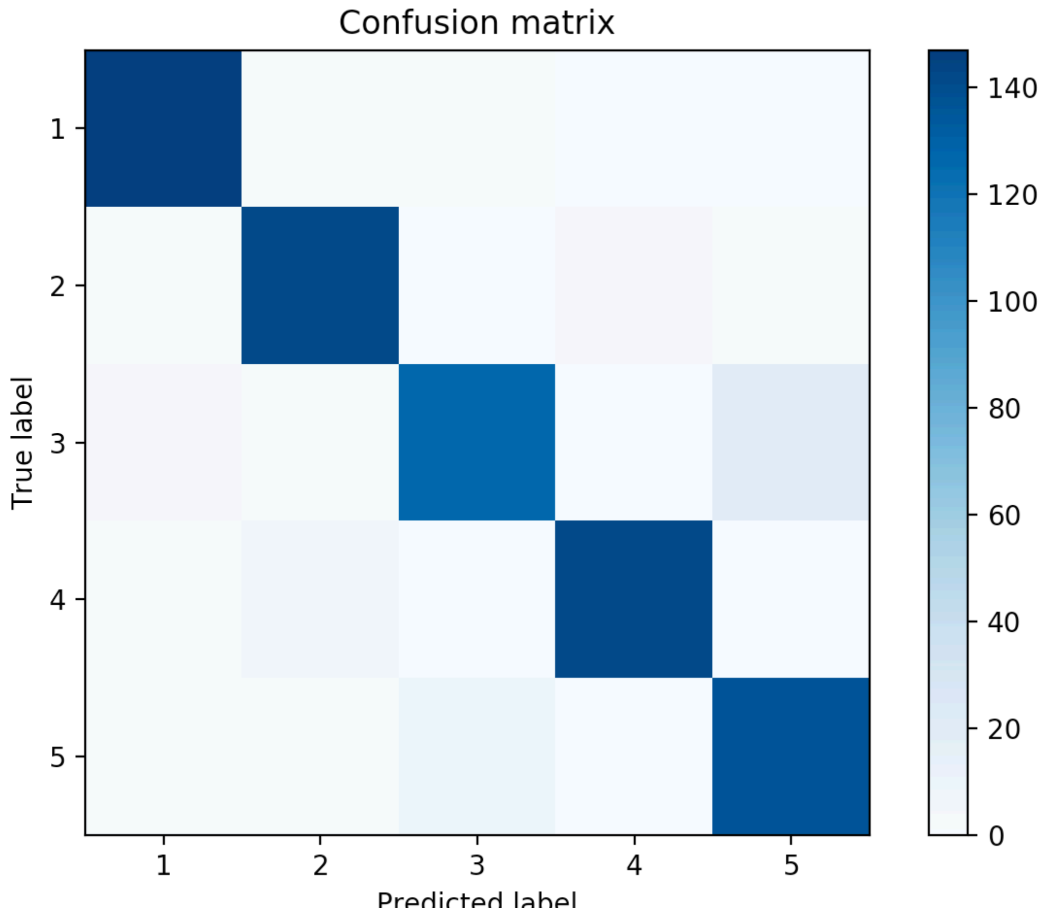
Images have a lot of features, which can be very time consumptive if you have to map to higher dimension, which other kernels do. The linear kernel gives an “accurate enough” result for image processing in less time than for example a rbf kernel.

Find the digit class for which you get the highest error rate.

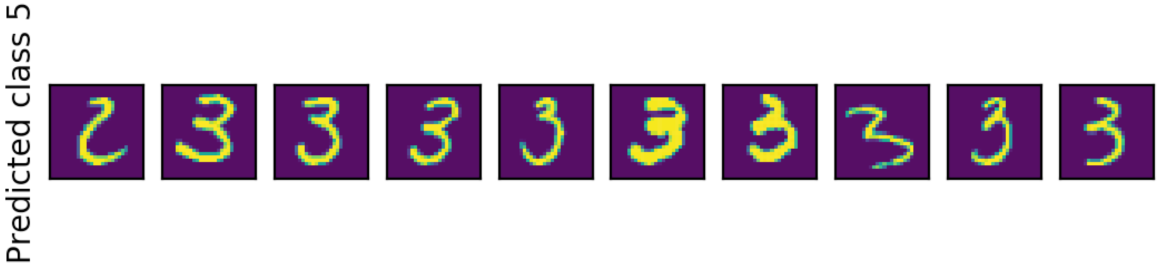
The most misclassified class was 5.

Include plots for ex_3_b of the confusion matrix and the first 10 images from the test set of the most misclassified digit.

Here is the confusion matrix for the given dataset with the linear kernel and OVR set as decision_function_shape:



First 10 images from the test set of the misclassified digit 5:



With the help of these two plots, discuss why the classifier is doing these mistakes.

As you can see in the confusion matrix, the real 3 is very often predicted as a 5, because the two digits have a lot of same parts, for example the second half circle of the 3 looks like the lower part of a 5. Vice versa there are some real 5’s predicted as 3.