2IIG0 - Homework 1 - Question 7

M.F.J. Moonen (1234115), Jin Ouyang (1608541), Bas Witters (1625187)

1 Subquestion A

For the linear classifier, the decision boundary does not fit the patterns visible in the given data because the data is not linearly separable. For both the red and the blue dots, a large amount of their respective points are not classified correctly.

The polynomial classifier is also not very good. We can see that it is trying to distinguish the data points in the center as red. But the blue data points are also included in the red-class areas. The data does not fit the form of a polynomial function, and so a polynomial classifier does not approach the true function that generated this dataset well.

The RBF classifier appears to fit the data quite well, and the shape given when plotting the resulting classifier seems to fit with the circular patterns we detect when looking at the data ourselves. This classifier also appears to be (close to) optimal, as it classifies most points correctly and most points that are misclassified are surrounded by points of the other type.

2 Subquestion B

The different combinations of C and γ can lead to different test accuracy. We find that a low γ value always causes the accuracy of the model to become significantly worse, whereas the value of C, other than for a specific combinations of C and γ, does not seem to influence the accuracy by a large margin.

To be more specific, when γ is lower than 0.01 (including 0.01), the test accuracy is under 0.6. When the γ is equal or larger than 0.1, the test accuracy will be higher when the parameter C is larger or equal to 1.0. When γ is larger than or equal to 1.0, the accuracy will be decent. We thus conclude that a large γ would be recommended, with preferably a corresponding value of C of 1.5.

We struggled to draw clear conclusions regarding under- or overfitting from only the heatmap, given that we need both train and test accuracy scores or the variances to determine this. Thus, we used both the heatmap and the train accuracy of each model in the GridSearchCV.

When comparing the train and the test accuracy for different configurations, we saw no large differences in either direction that could not be explained by differences between different train-test splits for different cross-validation runs. We conclude from this there is no clear underfitting or overfitting.

3 Optimal model accuracy & hyperparameters

- Test accuracy of best model: 0.7533
- Hyperparameters with best mean test accuracy: {'C': 2, 'gamma': 1}