Machine Learning

Lab 2: Support Vector Machines

Linus Groß
Daniel Mensah

Assignment 1

Move the clusters around and change their sizes to make it easier or harder for the classifier to find a decent boundary. Pay attention to when the optimizer (minimize function) is not able to find a solution at all.

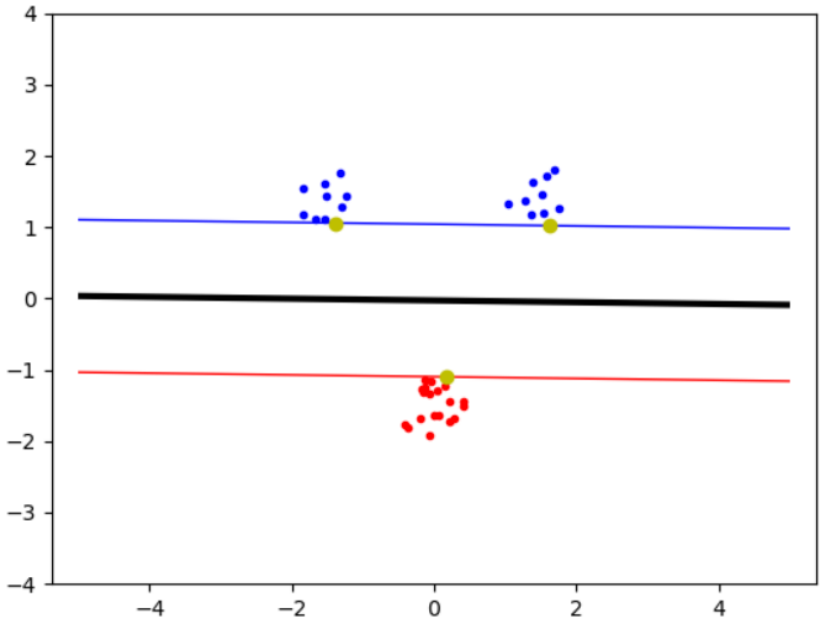

Figure 1: Easy classification because there is a clear boundary between the datasets. Furthermore there is a wide margin

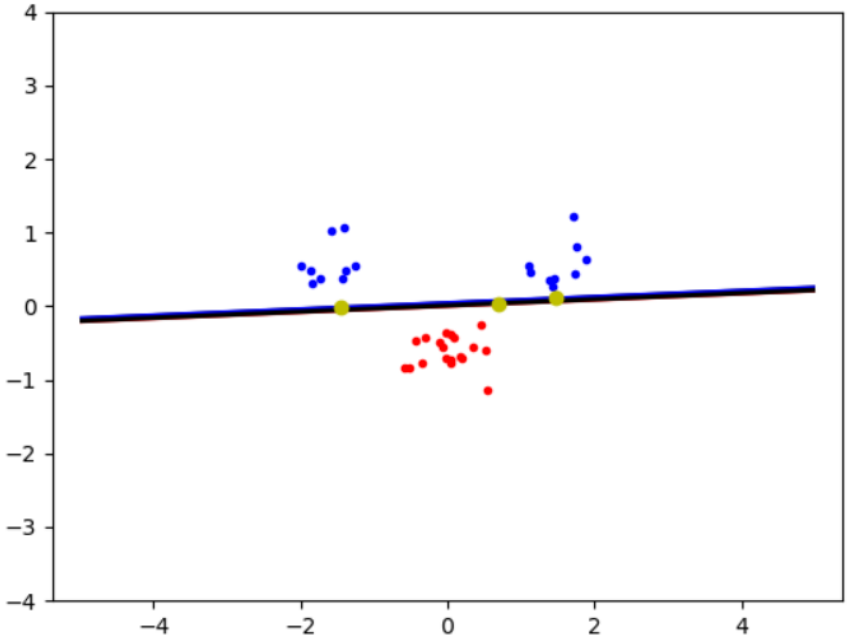

Figure 2: Higher variance and large clusters -> very small margin, because without slack the widest datapoints become the support vectors

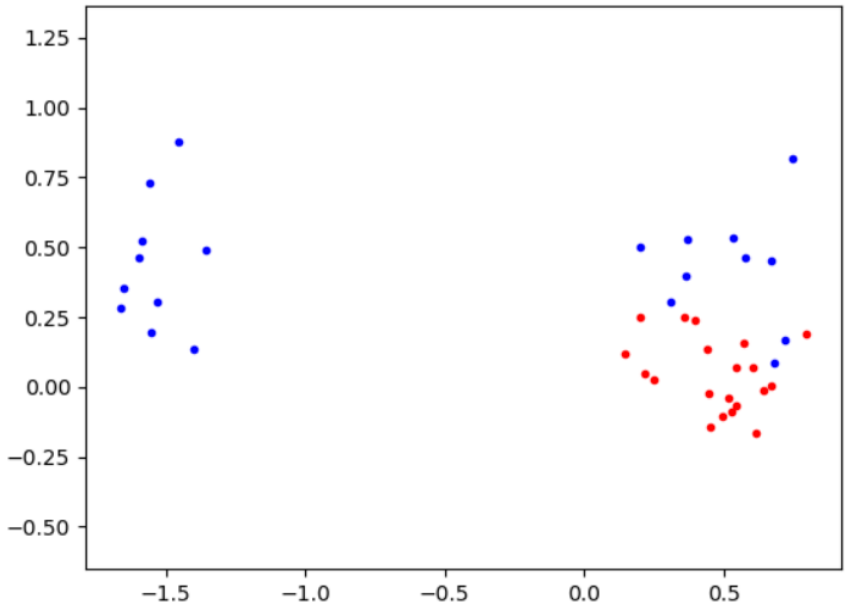

Figure 3: Overlapping datasets lead to no solution of the algorithm

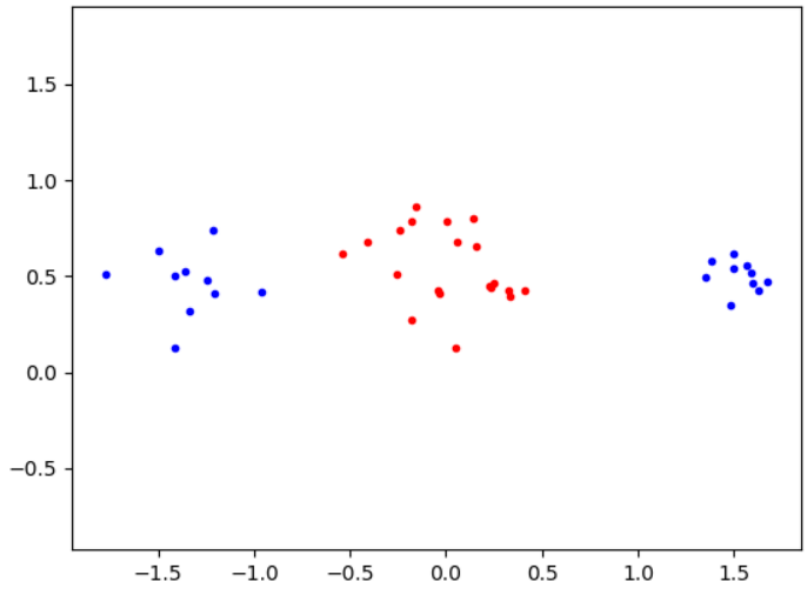

Figure 4: Data is not linearly separable, the algorithm finds no solution

The more datapoints are taken into count, the smaller the margin gets. This is because without any slack factor, the closest spread datapoints to the other classes will lead to the decision boundary. However, if a clear linear separation between the clusters can be found and they don't overlap, the algorithm founds a linear decision boundary.

If the clusters are arranged in a different way, so that there is no linear decision boundary, or the clusters overlap because of noise, the algorithm is not able to found a decision boundary.

Assignment 2

Implement the two non-linear kernels. You should be able to classify very hard data sets with these.

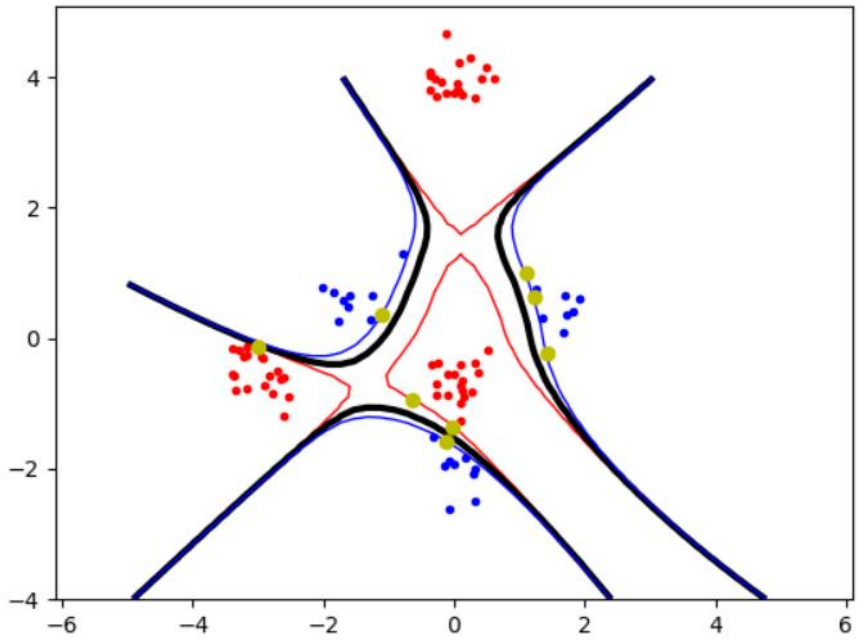

Figure 5: Polynomial kernel, p = 5

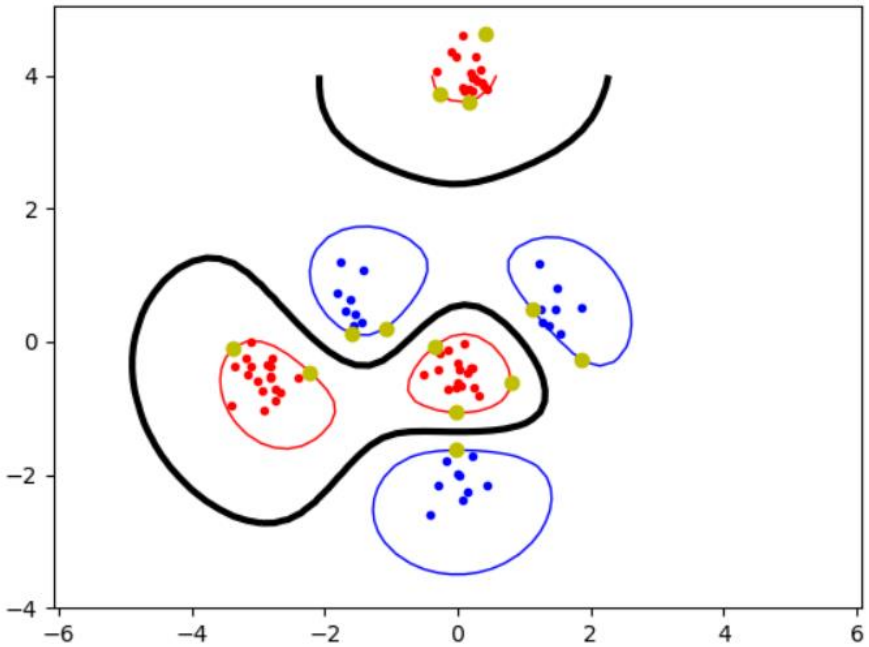

Figure 6: RBF kernel, σ= 1

Using the more complex kernels, the decision boundaries are not forced to be linear. This leads to curvy decision boundaries, which can classify more complex datasets, where there was no linear separation possible.

Assignment 3

The non-linear kernels have parameters; explore how they influence the decision boundary. Reason about this in terms of the bias-variance trade-off.

Polynomial kernel:

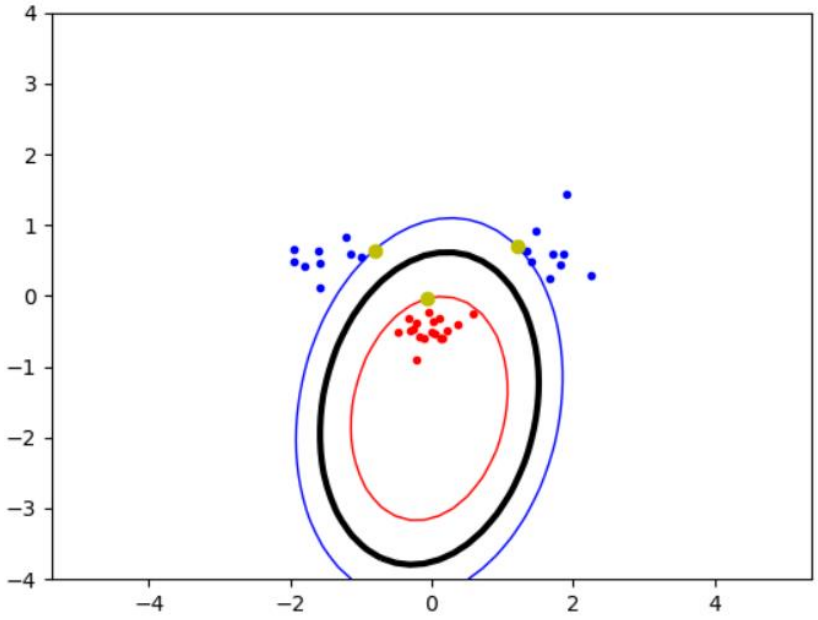

Figure 7: polynomial kernel, p = 2: round decision boundaries

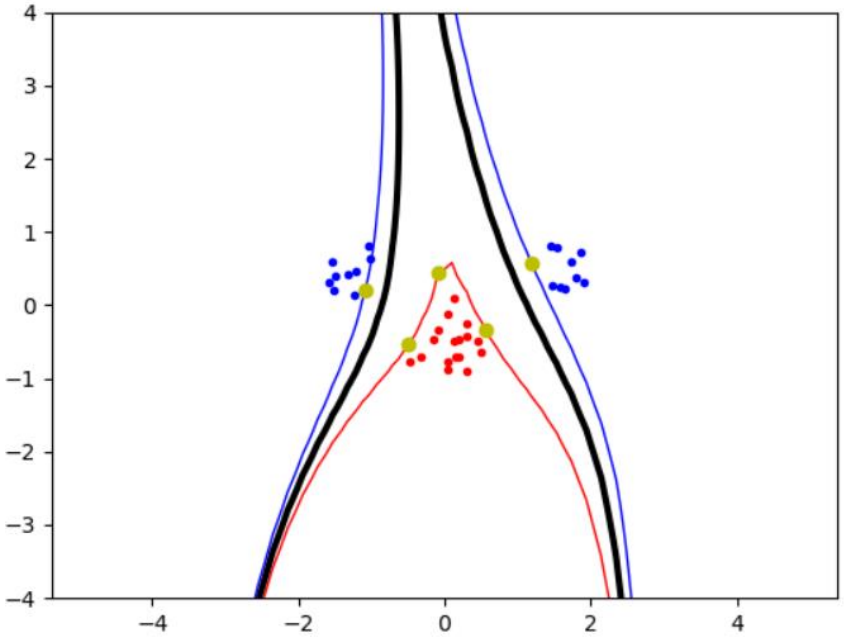

Figure 8: Polynomial kernel, p = 4: more complex shape of the decision boundary

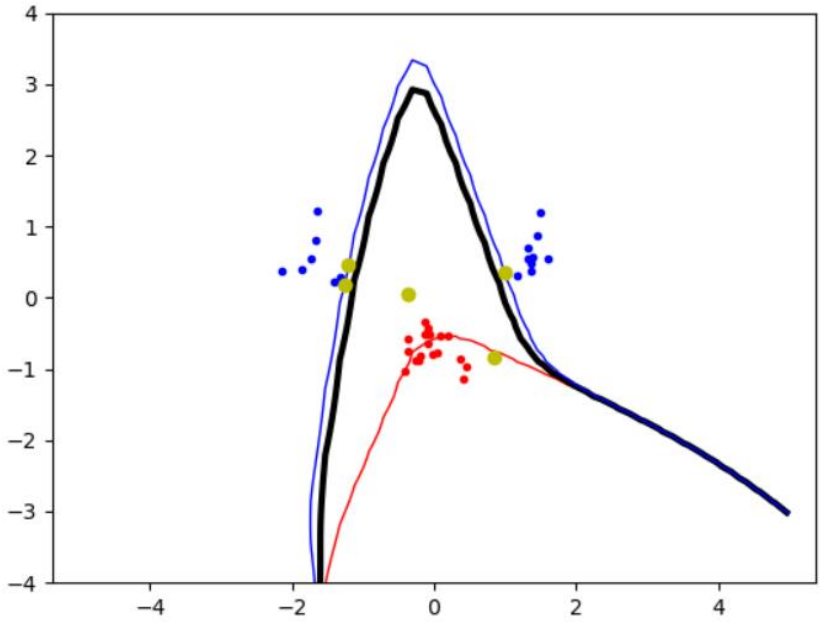

Figure 9: polynomial kernel, p = 10

The higher the order of the polynomial kernel is, the more complex shapes can be made as decision boundaries. This leads to a more complex distribution of the room and a higher variance of the classifier. A low order polynomial however leads to a higher bias because its smoother around the classes.

Radial Basis Function (RBF) kernel:

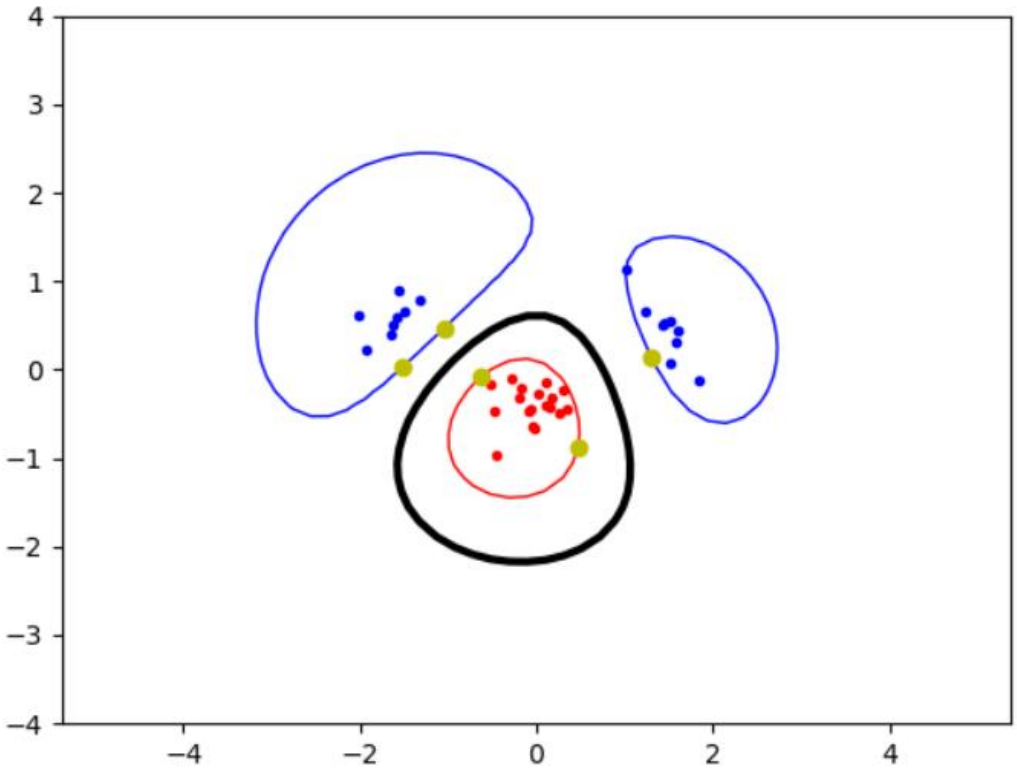

Figure 10: RBF-kernel, σ = 1: very good classification with round decision boundaries around the clusters, large margin

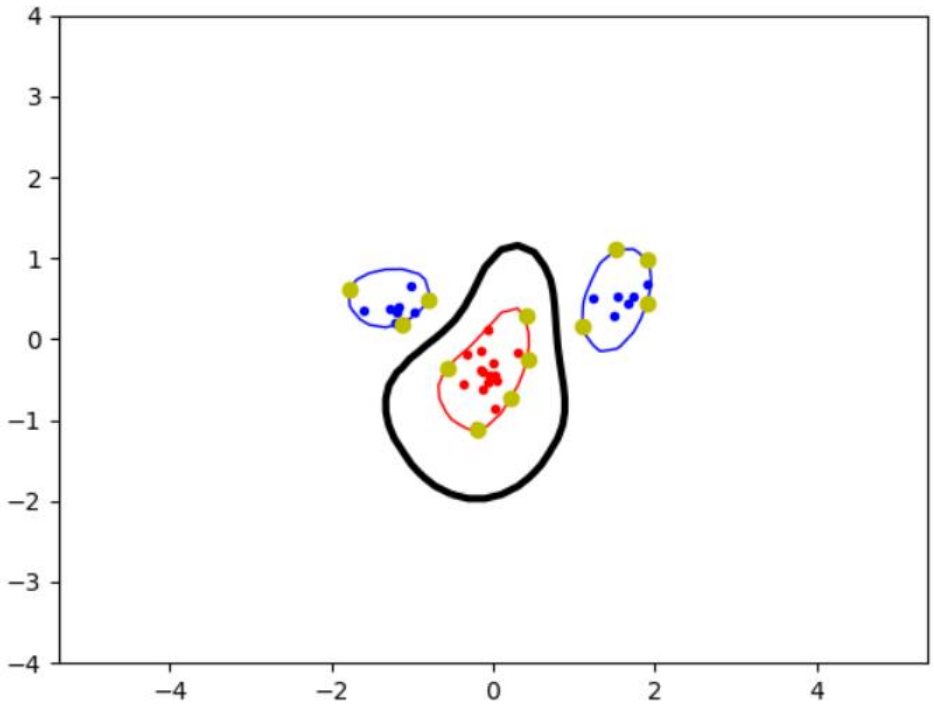

Figure 11: RBF-kernel, σ = 0.5: the decision boundaries get less smooth and more support-vectors are used, margin is around the clusters

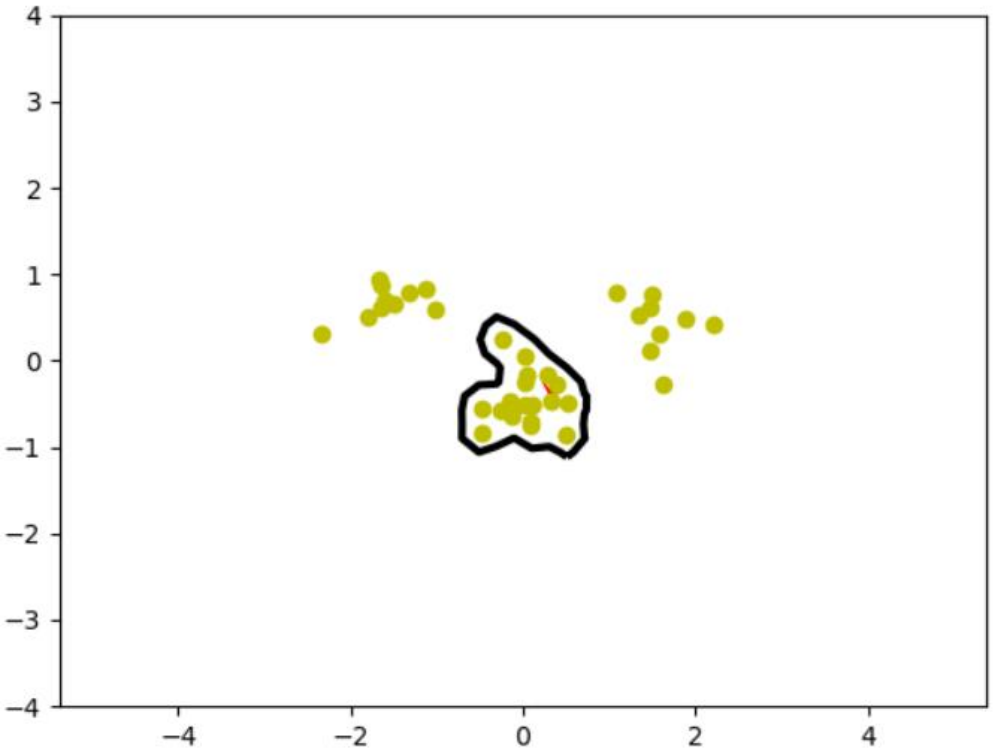

Figure 12: RBF-kernel, $\sigma = 0.1$: almost all datapoints are used for the decision boundary

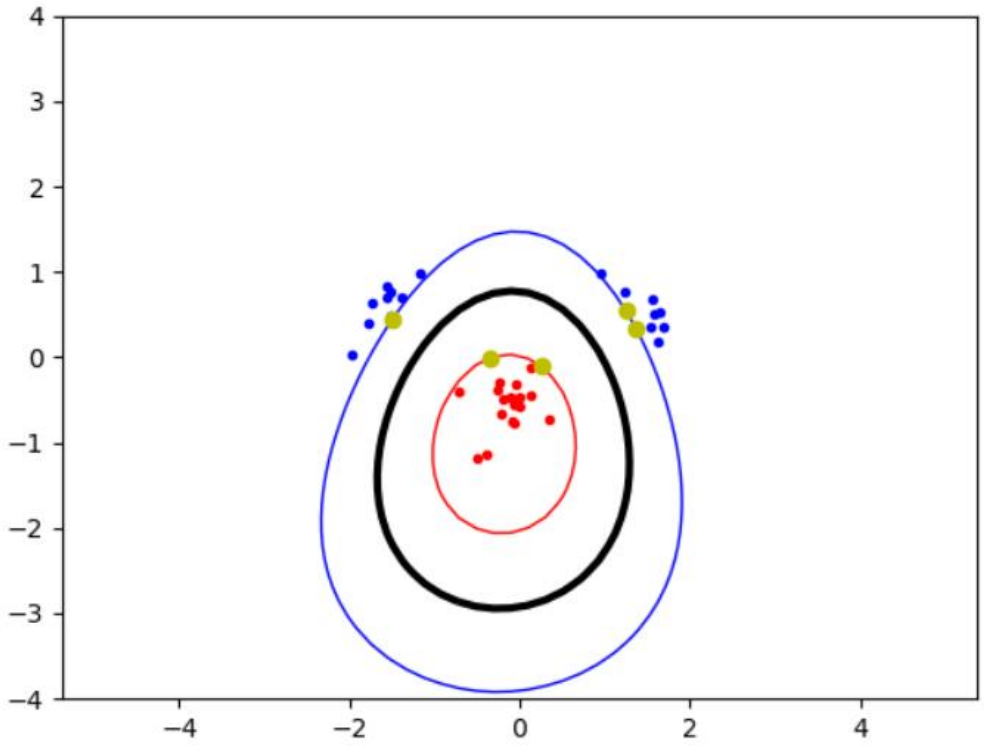

Figure 13: RBF-kernel, $\sigma = 2$: more clear decision boundaries which allow for a higher variance

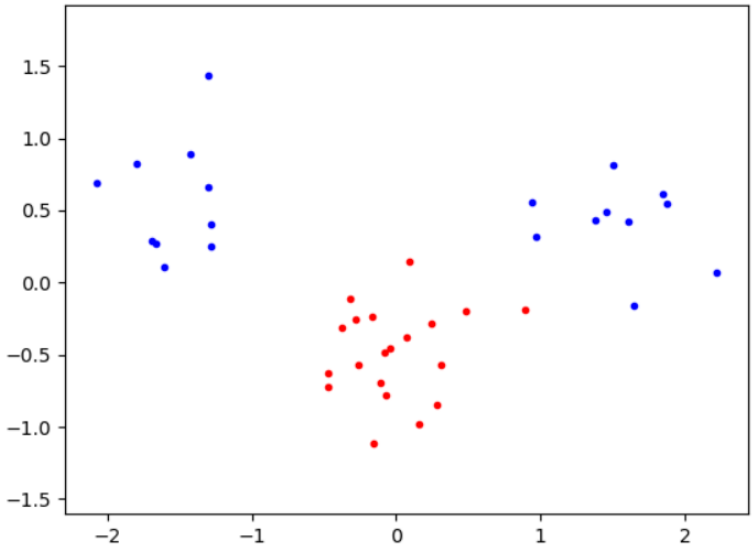

Figure 14: RBF-kernel, $\sigma = 10$: the optimizer is not able to find a solution without slack

For the RBF-kernel, the parameter σ determines the variance-bias tradeoff.

If you decrease σ, the decision boundary takes more datapoints into count which leads to smaller decision boundaries. The bias decreases, because the decision boundary just fits this one class. The variance however increases because the shape takes more variant forms.

Increasing σ leads to the opposite effect: The decision boundaries get wider and allow for a higher bias variance. Correspondingly the variance decreases. However, if the parameter is raised to high, the optimizer is not able to find a solution, if no slack is allowed. This is because the allowed bias is too big and would lead to datapoints of the training set within the margin.

Assignment 4

Explore the role of the slack parameter C. What happens for very large/small values?

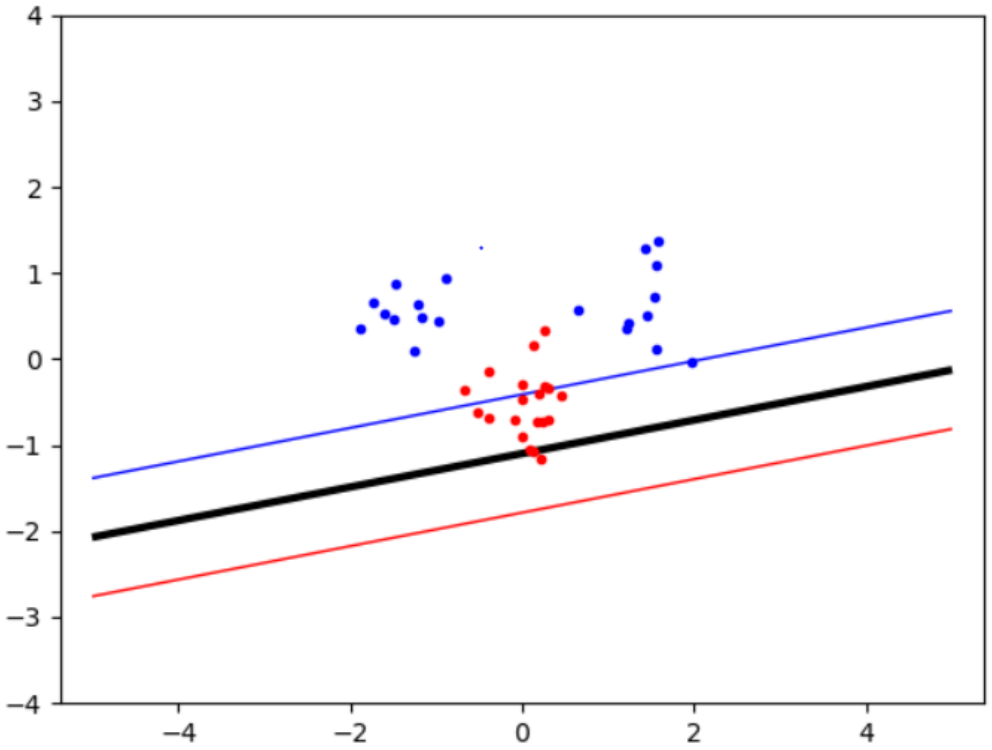

Figure 15: C = None, without the slack-parameter, no linear solution can be found

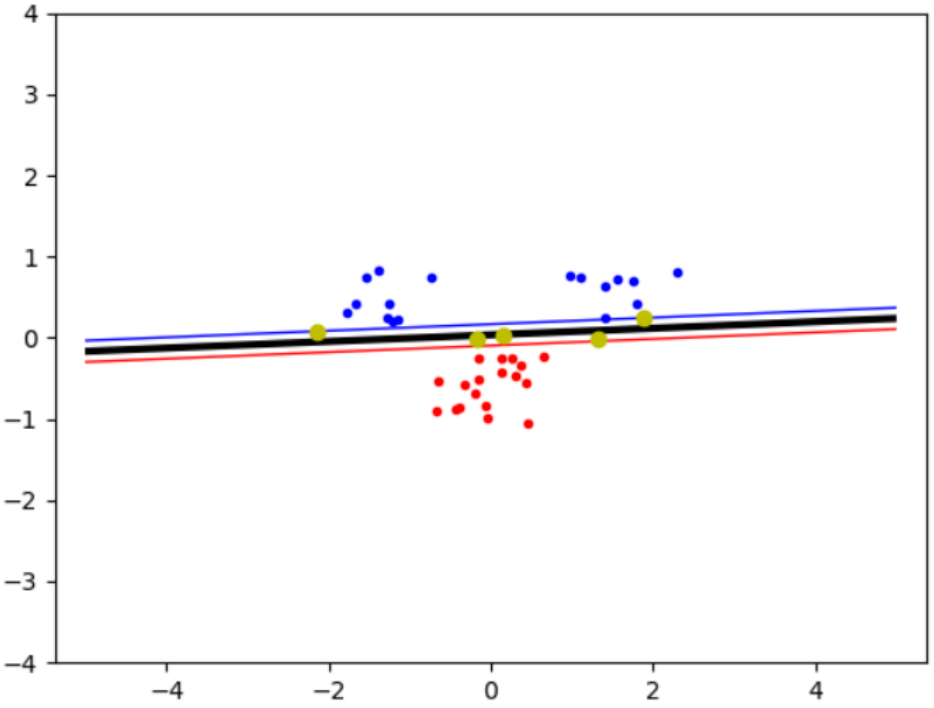

Figure 16: C = 100: Allows for a few datapoints to be into the margin-zone

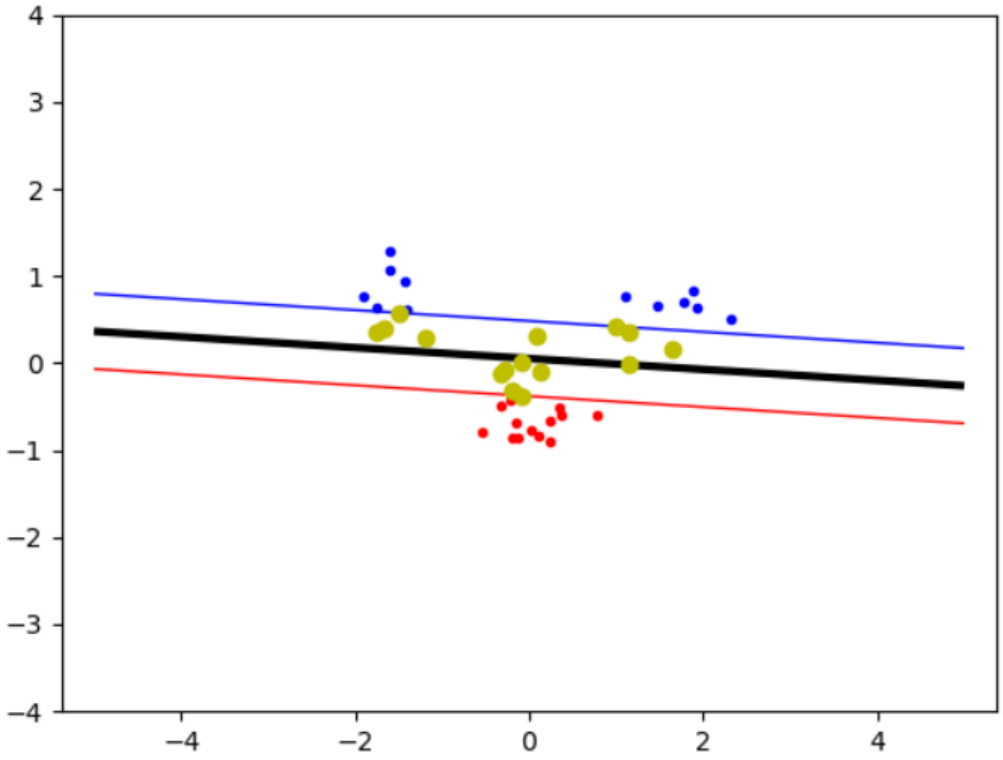

Figure 17: C = 1: lower slack parameter leads to a more datapoints within the wider margin

Without any slack parameter, no test data is allowed to be in the margin, only support vectors sit on the edge of the playroom. A noisy data however can lead to wider spread datapoints, which then prevent a decent decision boundary. With the introduction of the slack parameter, datapoints are allowed to be within the margin so that noisy data with a few overlapping points can be classified by a decision boundary. The slack parameter C must be chosen by hand.
Lowering the slack parameter leads to a higher tolerance of datapoints staying in the margin. This in turn leads to a wider margin. A too wide margin however can lead to an unclear decision boundary.

Assignment 5

Imagine that you are given data that is not easily separable. When should you opt for more slack rather than going for a more complex model (kernel) and vice versa?

If the data is very noisy, you should go for more slack. This allows for more tolerance for datapoints, which are more widely spread and can overlap to other classes.
When however, the classes are separated and form complex decision boundaries, more complex kernels should be used.