SVM

## What to explore

1. Move clusters around
2. Vary slack variable, large to small
3. Explore different cluster combinations and kernel functions
4. Reason why certain kernel functions are more suitable

RBF works like a k nearest neighbour

The bigger the difference the smaller the kernel value and its influence or relationship between the two points.

Gamma determines the size of the influence, higher gamma smaller influence

<https://www.quora.com/What-is-the-purpose-for-using-slack-variable-in-SVM> On slack variable

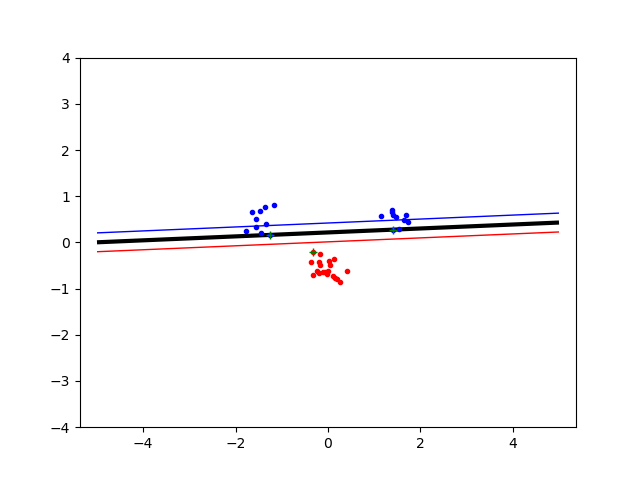
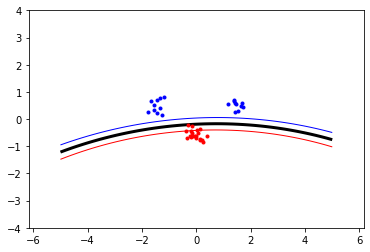
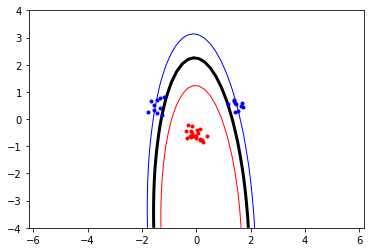


Figure 1. For all slack = 0, Left: Poly, r = 1, p = 2, Right: RBF, sigma = 10,

Bottom: Linear

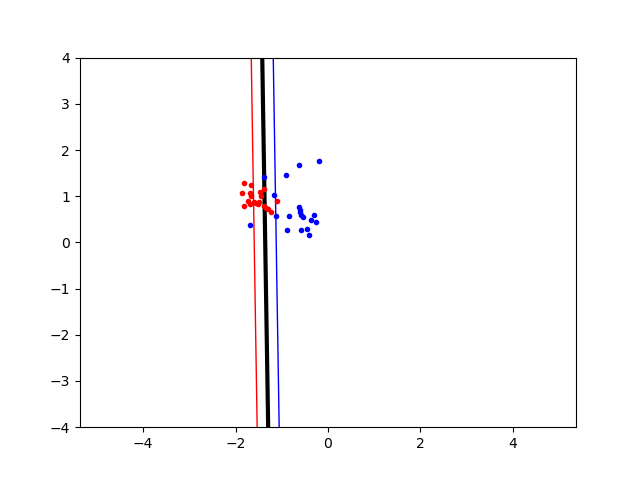
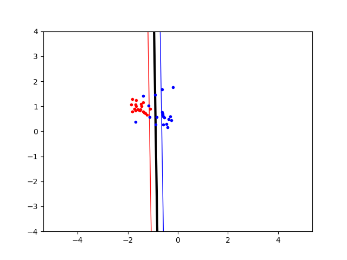
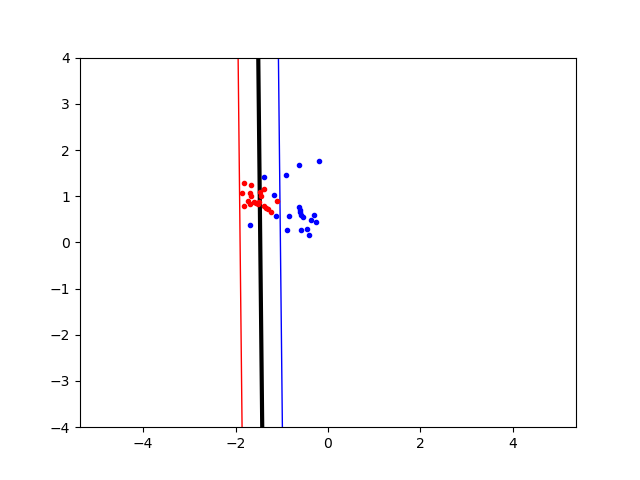
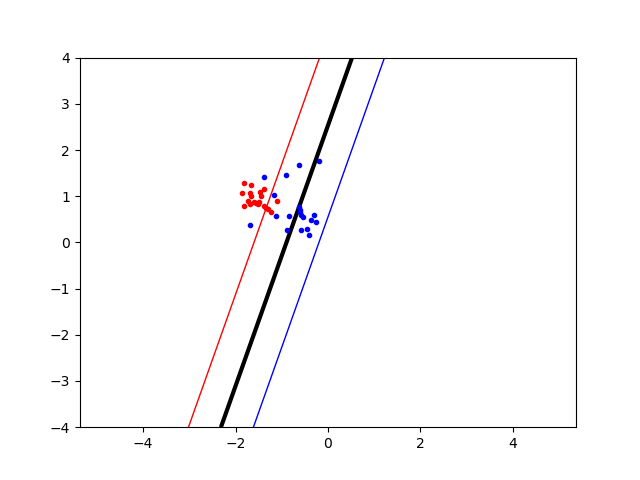
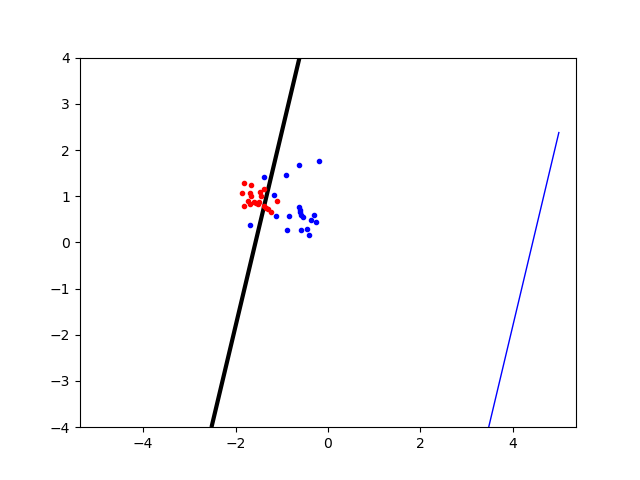


Figure 2. Linear kernel for dataset where some blue points are scattered in red cluster. Plotted for slack variable with a value of [0.01, 0.1, 1, 10, 500] from top to bottom.

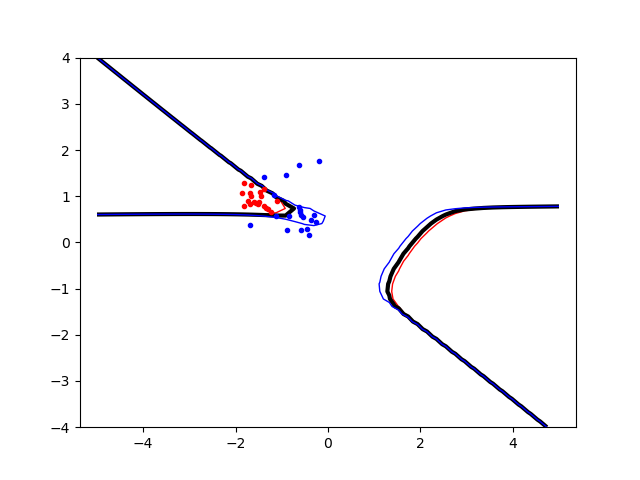
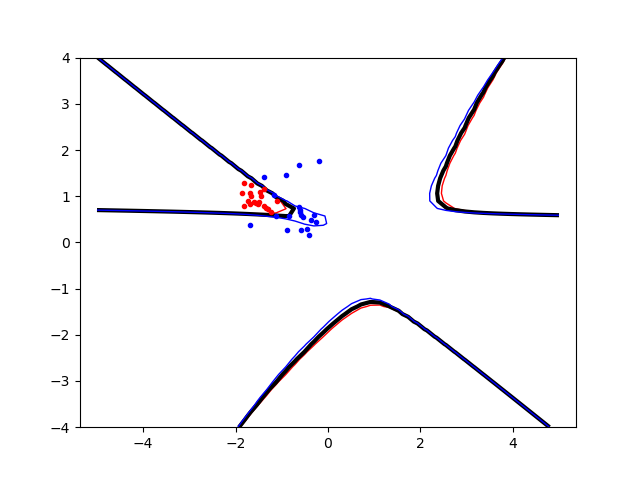
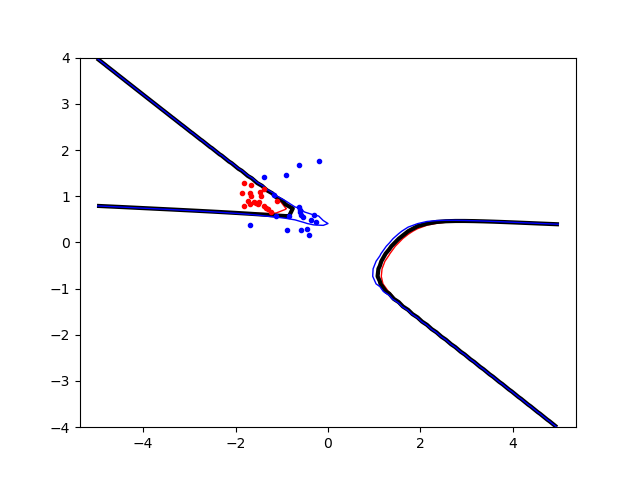
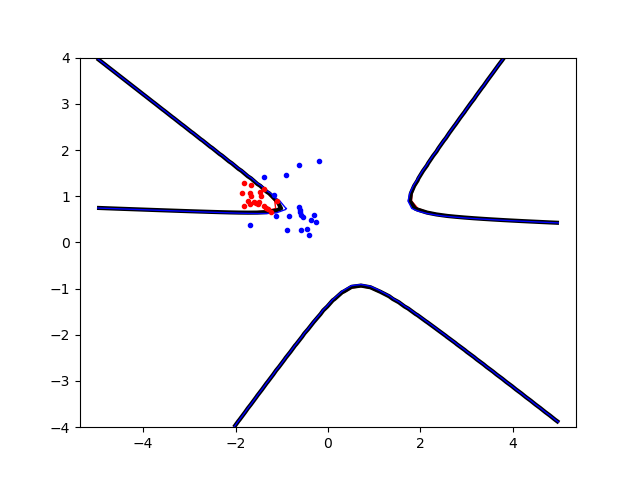


Figure 3. Same dataset as Figure 2. But here the Poly kernel is used. Kernel parameters R = 1 and P = [3, 4, 5, 6] from top to bottom.

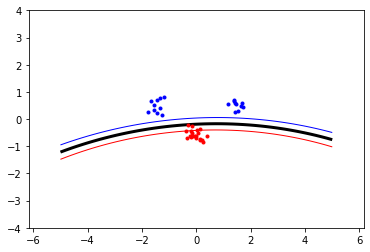
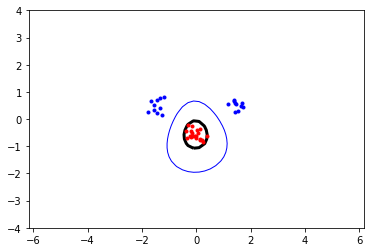
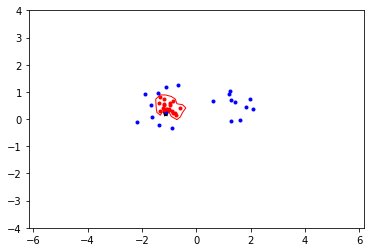


Figure 4. Original dataset with the RBF kernel. Sigma values = [0.1, 1, 10] from top to bottom.

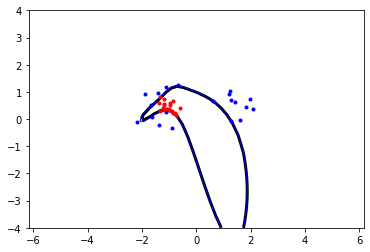
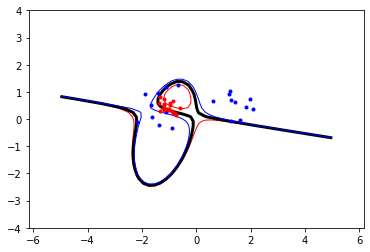
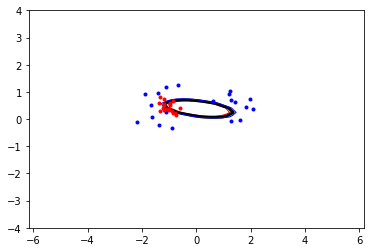


Figure 5. Poly kernel on slightly scattered dataset where the red class is caught in the middle of the blue class. Kernel parameters R = 10 and p = [2, 3, 5] from top to bottom.

### 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.

Using a more complex kernel would arguably decrease bias, it fits better to the dataset. But since we do not know the true distribution a point could appear anywhere and variance would thus become bigger with a more complex model.

### 4. Explore the role of the slack parameter C. What happens for very

### large/small values?

Slack can be viewed as a penalty on the error. A smaller slack value C would mean you’ll allow the SVM to misclassify more datapoints than if you’d have a larger C. A smaller C also seems to increase the amount of support vectors.

### 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?

In Figure 2. and Figure 3. we can see the effect of using a linear kernel with more or less slack versus a polynomial kernel with higher or lower exponent. It seems that for this data it would be better with a simpler kernel and more slack.