Support Vector Machines

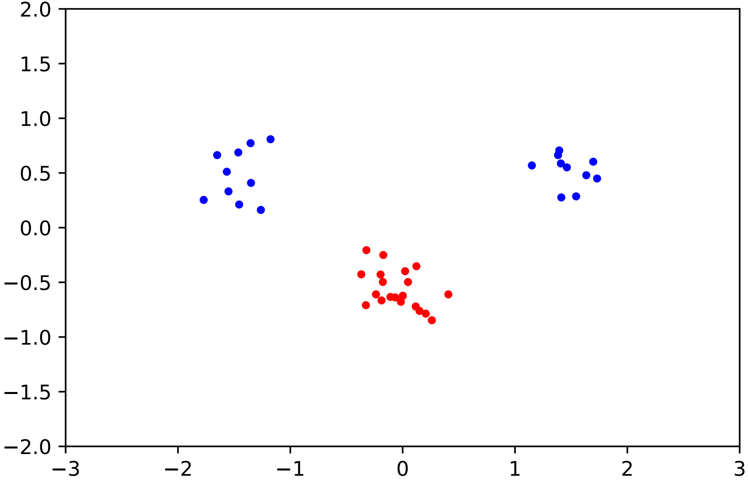
Marina Herrera, Thi Thuy Nga Nguyen

Assignment 2 - DD2421 Machine Learning
KTH Royal Institute of Technology, Sweden

February 20th, 2019

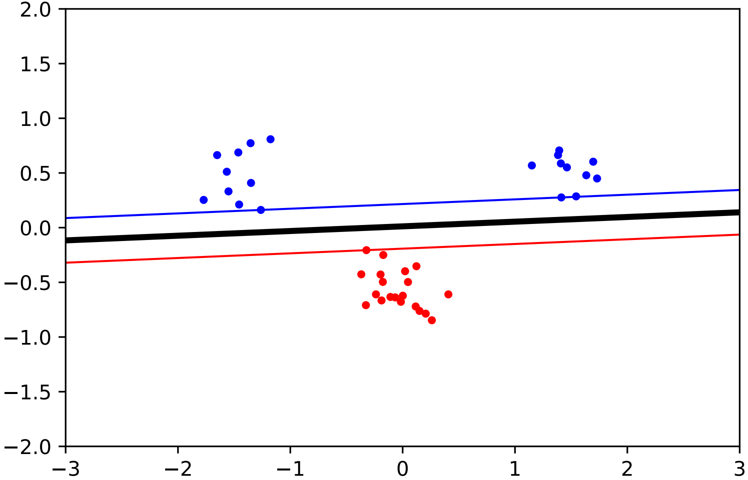
Linear kernel

Maximaize the margin (or distance to any datapoint).



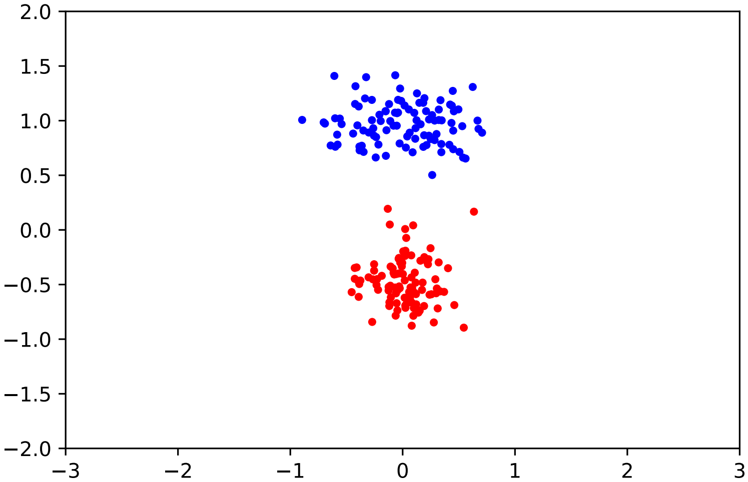
Linear kernel

Maximaize the margin (or distance to any datapoint).



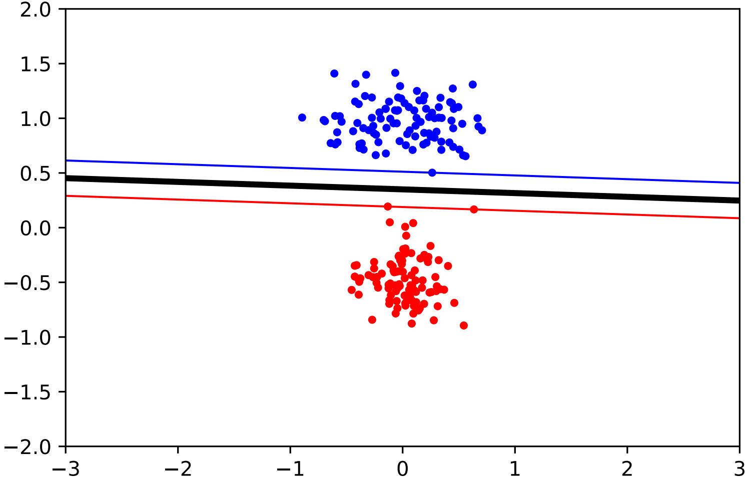
Linear kernel

Take longer time for large datasets.



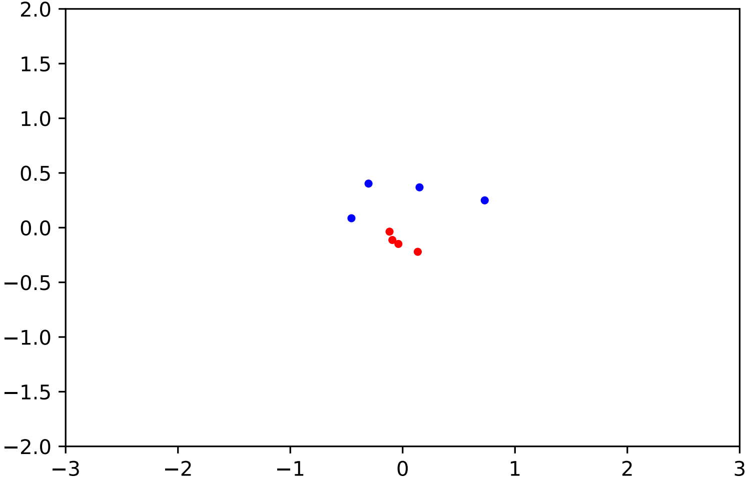
Linear kernel

Take longer time for large datasets.



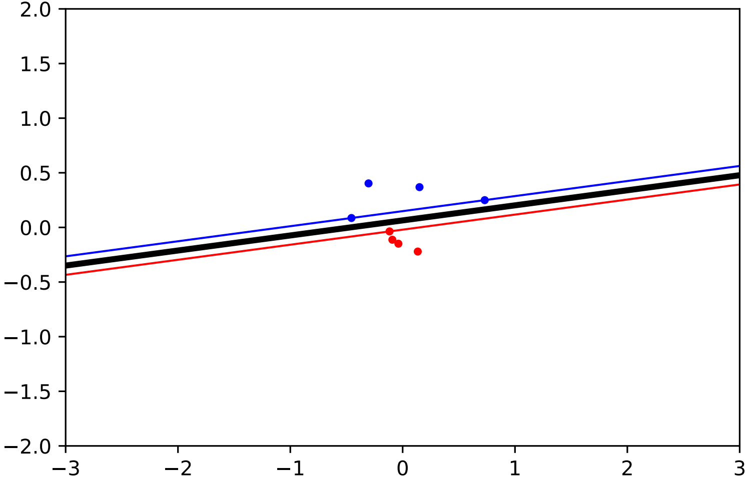
Linear kernel

Very good for small sizes samples.



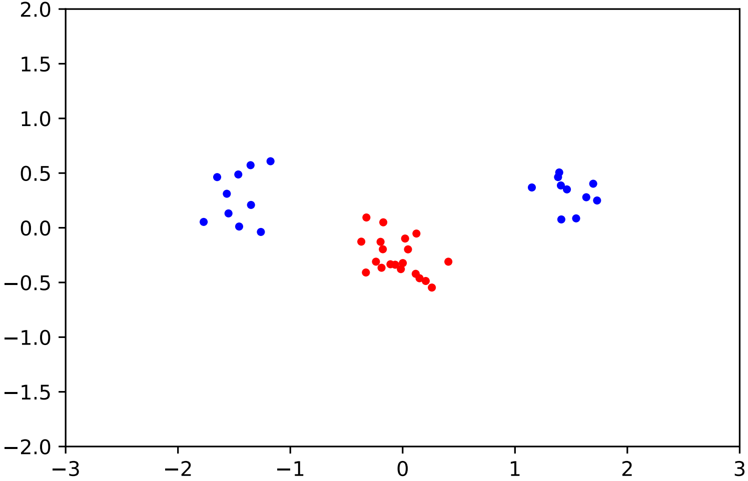
Linear kernel

Very good small sizes samples.



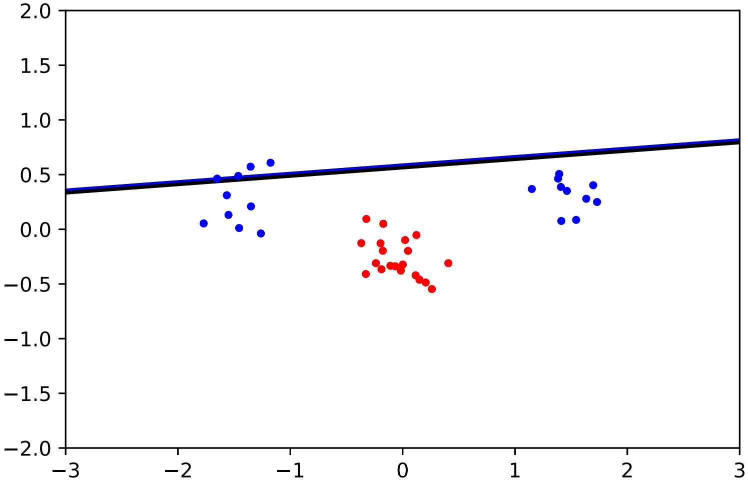
Linear kernel

No solution.



Linear kernel

No solution.



Non-linear kernels

- Polynomial kernels,

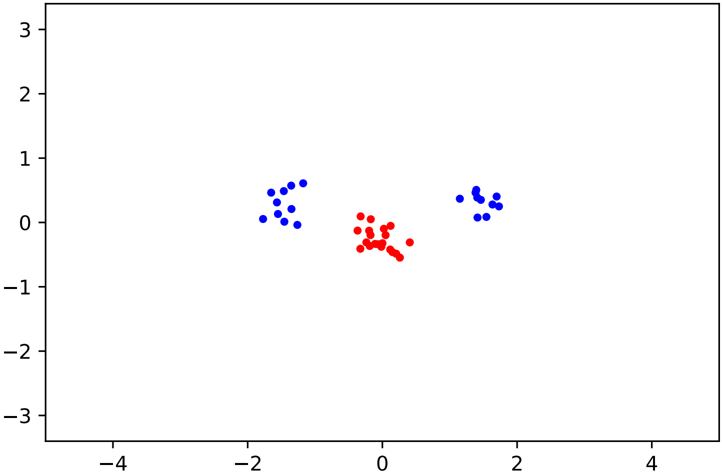
$$\mathcal{K}(\vec{x}, \vec{y}) = \left(\vec{x}^T, \vec{y} + 1\right)^p .$$

- Radial Basis Function (RBF) kernels,

$$\mathcal{K}(\vec{x}, \vec{y}) = e^{-\frac{||\vec{x}-\vec{y}||^2}{2\sigma^2}} .$$

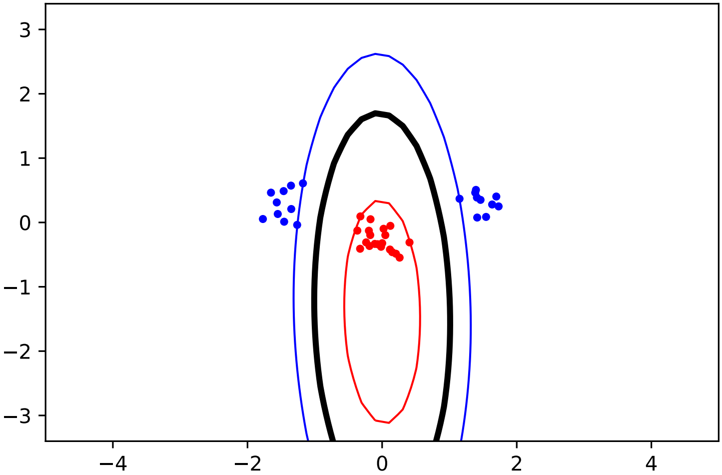
Non-linear kernels

Polynomial kernels, $p = 2$.



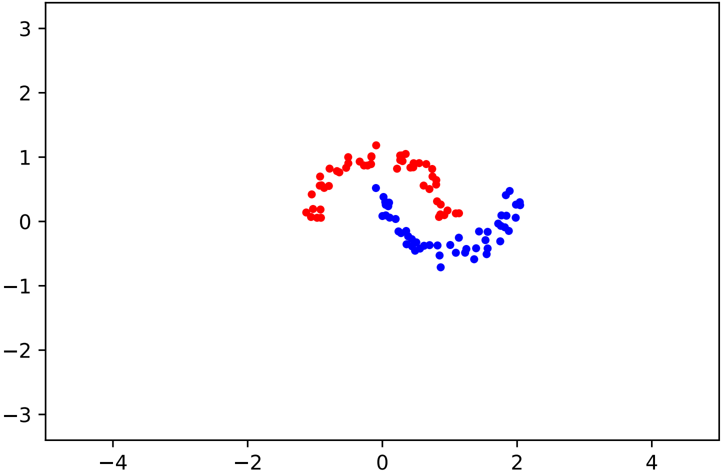
Non-linear kernels

Polynomial kernels, $p = 2$.



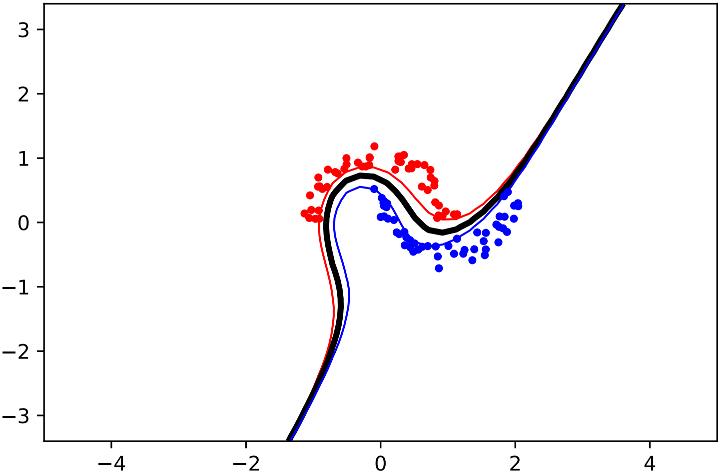
Non-linear kernels

Polynomial kernels, $p = 3$.



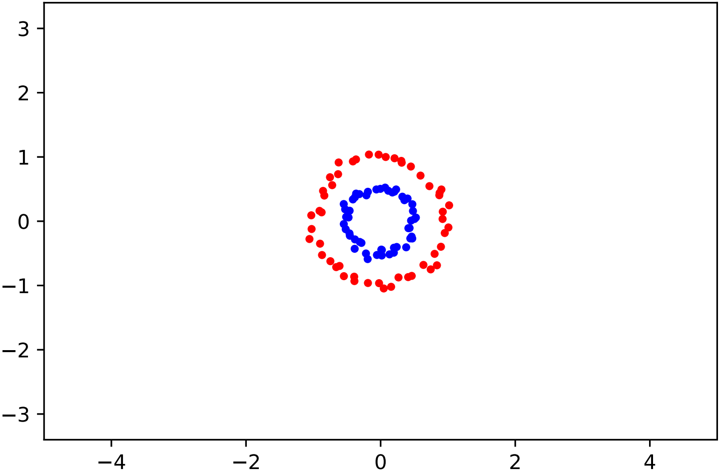
Non-linear kernels

Polynomial kernels, $p = 3$.



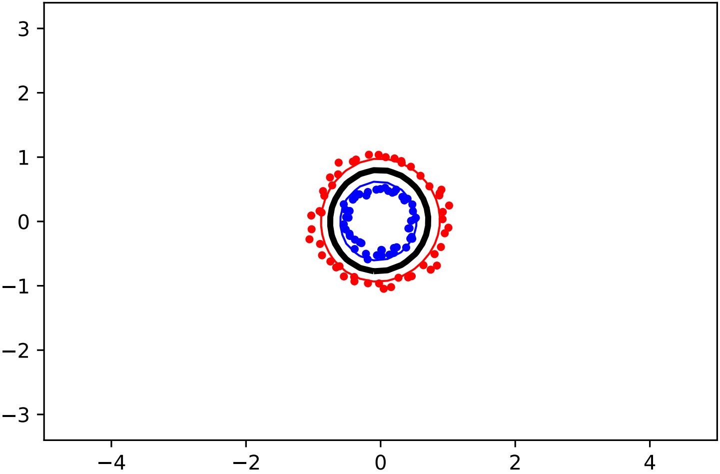
Non-linear kernels

Radial Basis Function (RBF) kernels, $\sigma = 1$,



Non-linear kernels

Radial Basis Function (RBF) kernels, $\sigma = 1$.



Bias-variance trade-off and the parameter of the kernels

Polynomial kernels:

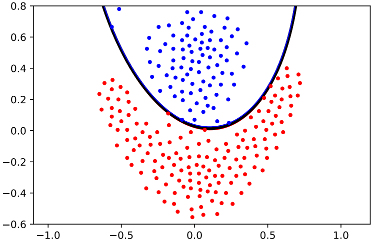
A large p corresponds to more complex decision boundary which implies high variance and low bias and vise versa.

RBF kernels:

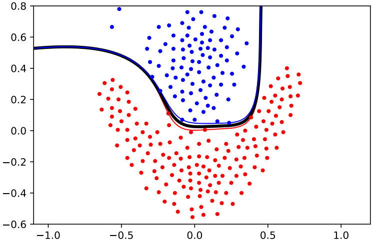
A large σ corresponds to a small value of kernel function that means the support vector does not have much influence on the classification. It allows more complex decision boundary but it faces on overfitting. Therefore, a large σ leads to high bias and low variance model.
A small σ implies that the SV has larger influence on the classifying. The decision boundary thus becomes simpler that leads to high bias, low variance model.

Bias-variance trade-off and the parameter of the kernels

Polynomial kernels



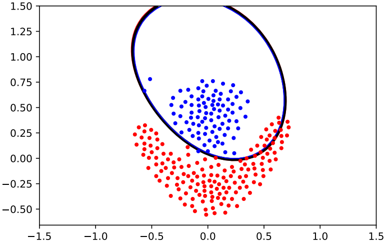
$p = 2$



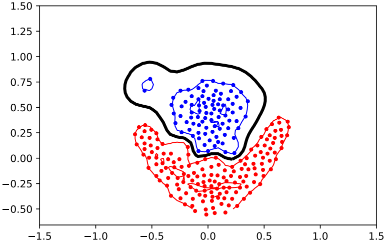
$p = 20$

Bias-variance trade-off and the parameter of the kernels

RBF kernels.



$\sigma^2 = 1$



$\sigma^2 = 0.1$

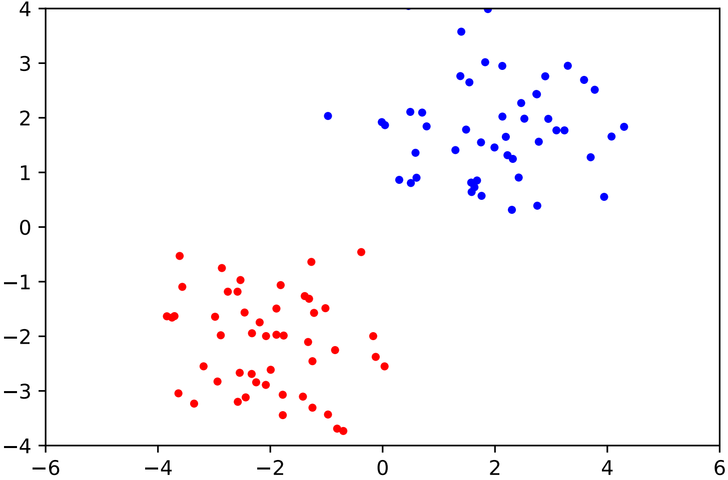
Slack parameter C

Q4: Explore the role of the slack parameter C. What happens for very large/small values?

A: The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable.
A large C leads to low bias and high variance model.

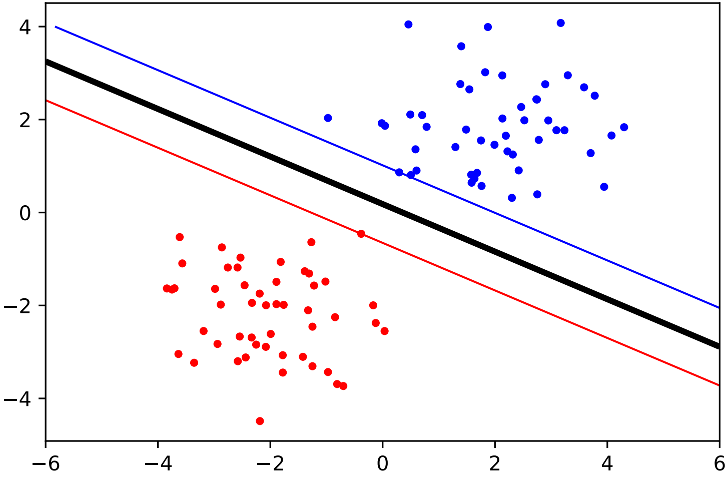
Slack parameter C

Linear classifier



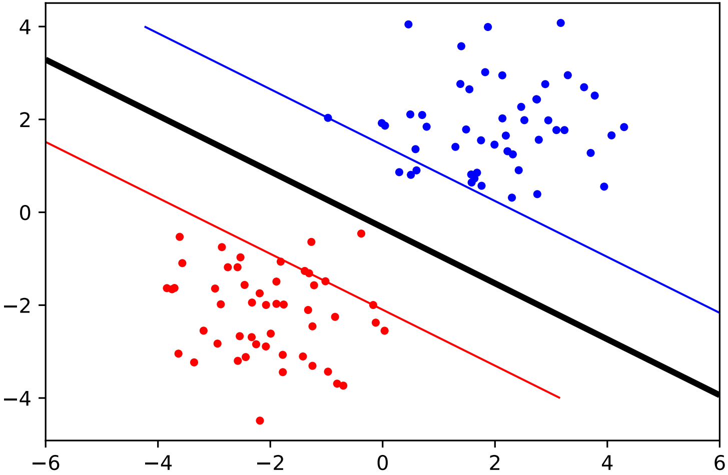
Slack parameter C

Linear classifier without C.



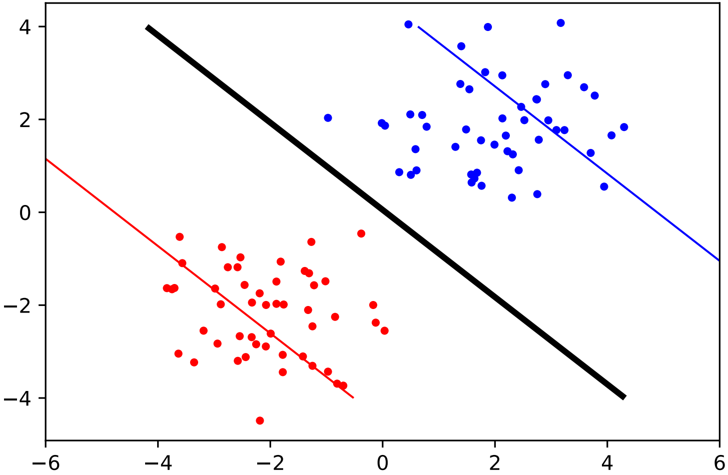
Slack parameter C

Linear classifier with $C = 0.1$.



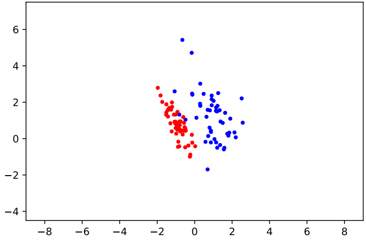
Slack parameter C

Linear classifier with $C = 0.002$.



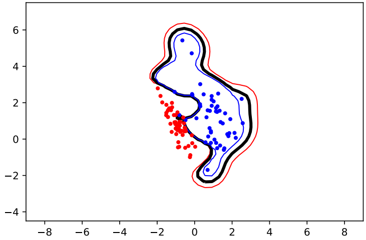
More slack or more complex model

Dataset has much noise

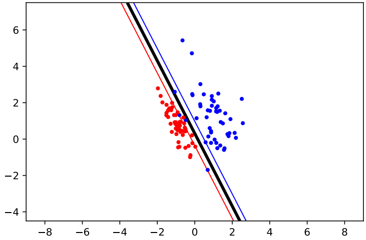


More slack or more complex model

Dataset has much noise $\rightarrow$ choose a good slack.



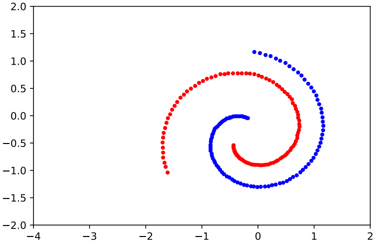
RBF kernel with $\sigma = 0.7$.



Linear kernel with $C = 10$

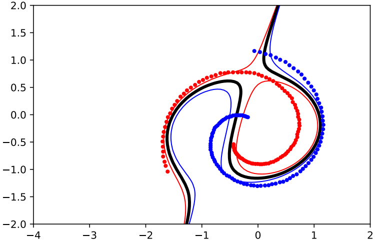
More slack or more complex model

Dataset has no noise with complex shape

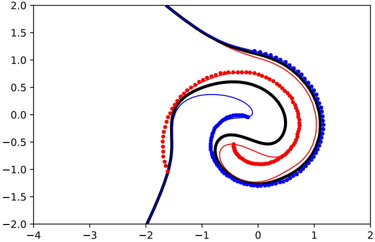


More slack or more complex model

Dataset has no noise with complex shape $\rightarrow$ choose a complex model.



Polynomial kernel with $p = 3, C = 10$.



Polynomial kernel with $p = 10$.