MACHINE LEARNING LAB 2

SUPPORT VECTOR MACHINES

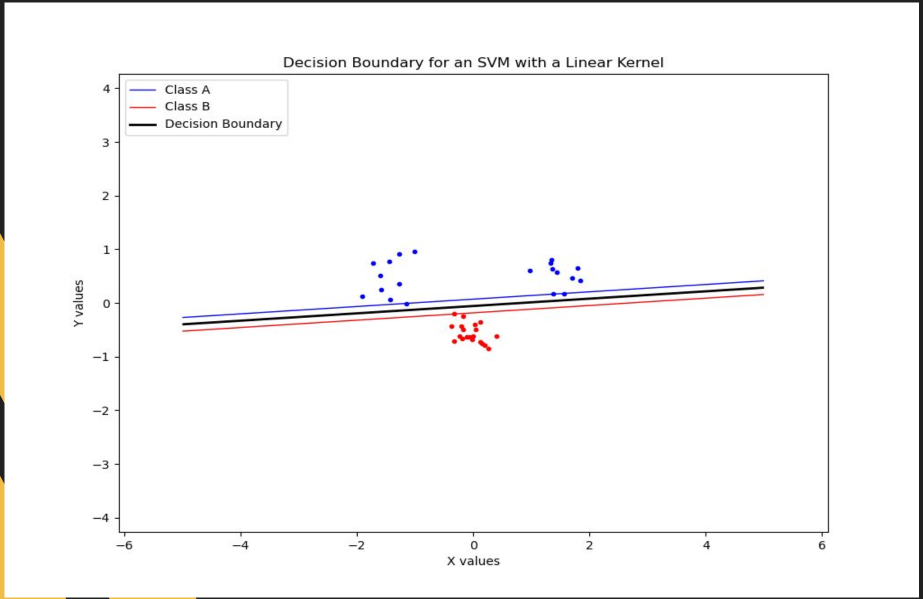
David O'Leary & Cillian Smith

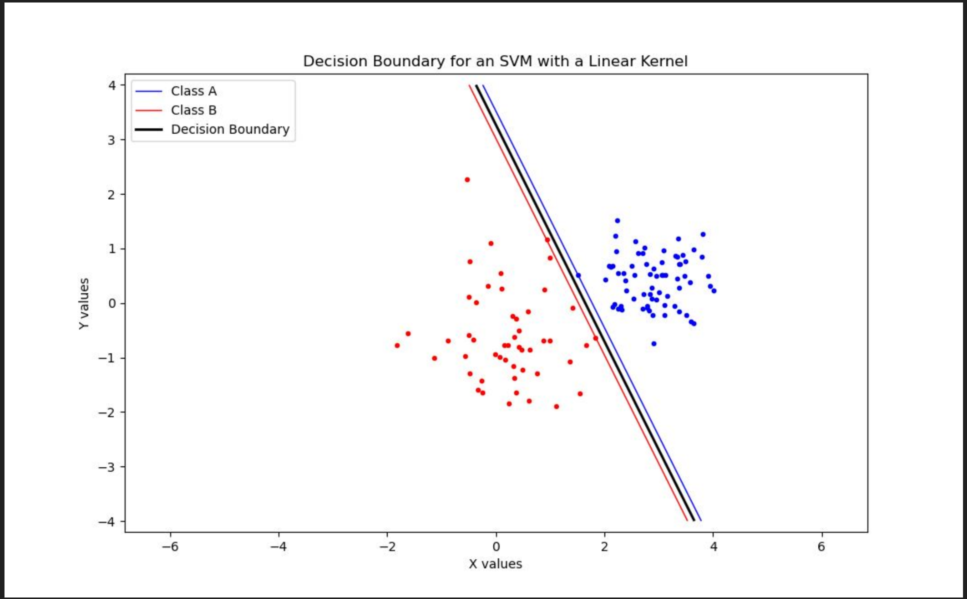
LINEAR KERNELS

- Support Vector Machines (SVMs) are used for binary classification problems.
- They are robust to noise as they attempt to maximize the margins between classes. As such, they are less influenced by individual data points.
- Linear kernels work well when classes of data are linearly separable in the feature space.
- The optimiser will fail when the classes are no longer linearly separable, i.e the classes being able to overlap.
  - This can be avoided by adjusting the slack.

$$K(X,Y) = X^T \cdot Y$$

LINEAR KERNELS





SVMs for two different datasets using a linear kernel

NON-LINEAR KERNELS

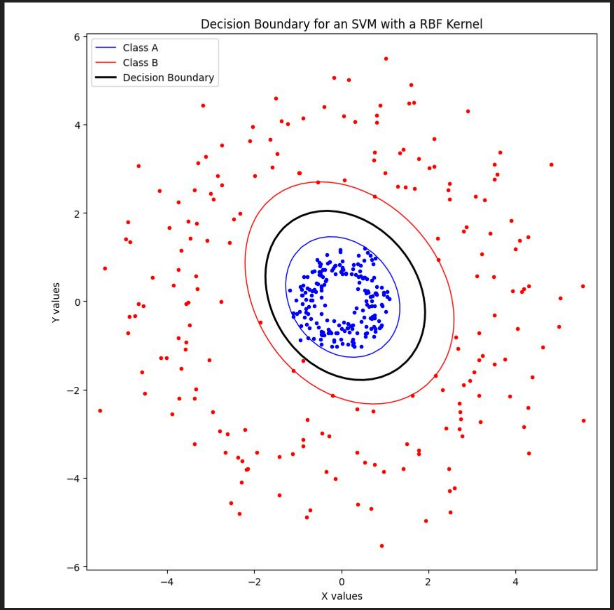
Polynomial Kernel

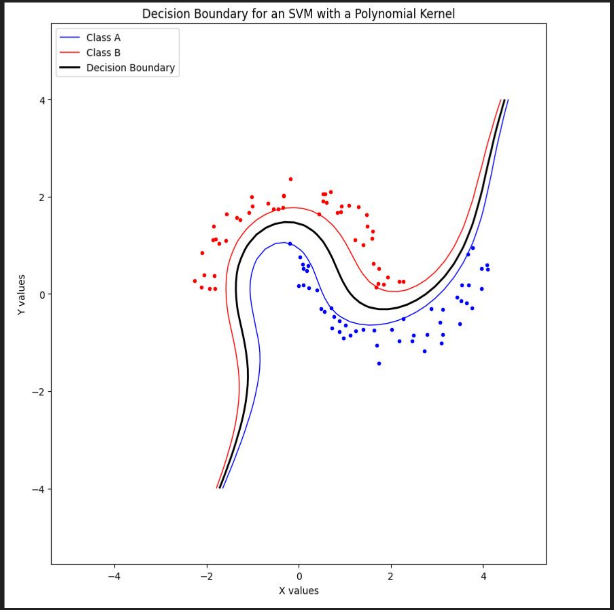
$$K(X, Y) = (1 + X^T Y)^p$$

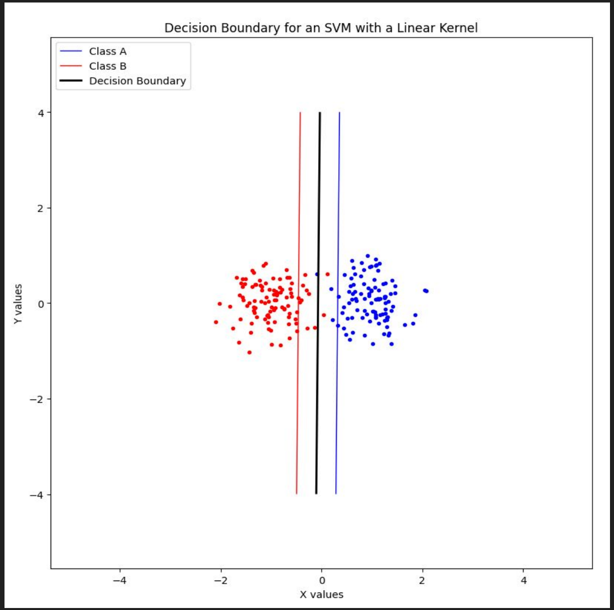
RBF Kernel

$$K(X, Y) = \frac{||X - Y||^2}{2\sigma^2}$$

CHOOSING A KERNEL







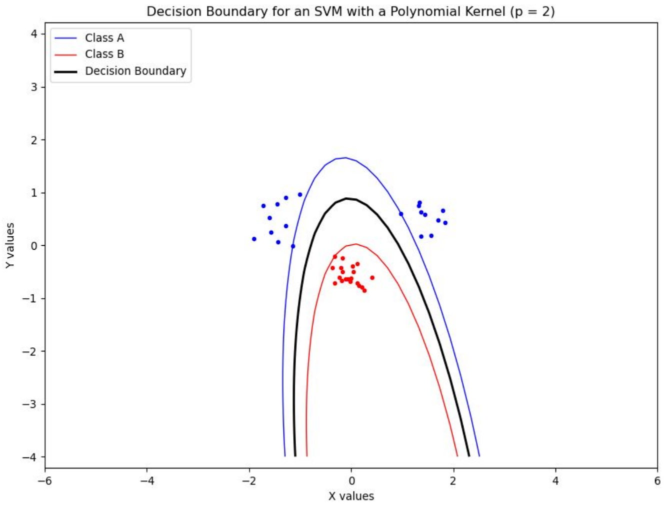
POLYNOMIAL KERNELS

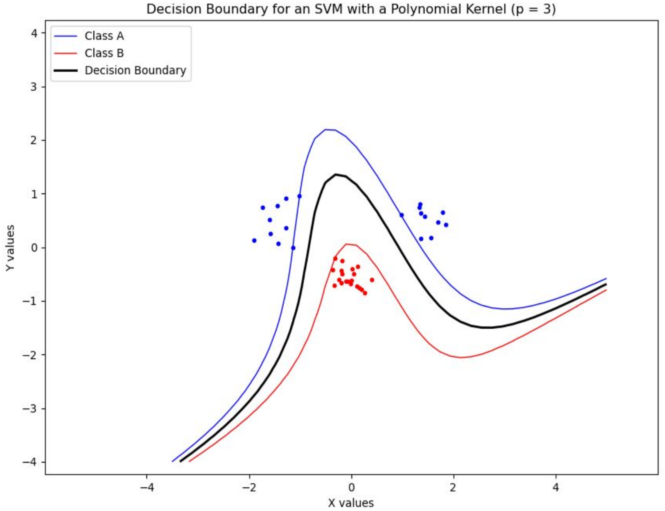
$$K(X, Y) = (1 + X^T Y)^p$$

Polynomial kernels are parameterised by P, which represents the degree of the polynomial curve used for mapping the input data.

Bias-variance trade-off:

- Increasing P :
  - Increases the variance, as a higher degree polynomial can capture more complex input data. This can also decrease the bias
- Decreasing P :
  - Decreases the variance, a degree of 2 polynomial curve would be unable to capture a complex underlying distribution, increasing the bias.





RBF KERNELS

$$K(X,Y) = \frac{||X - Y||^2}{2\sigma^2}$$

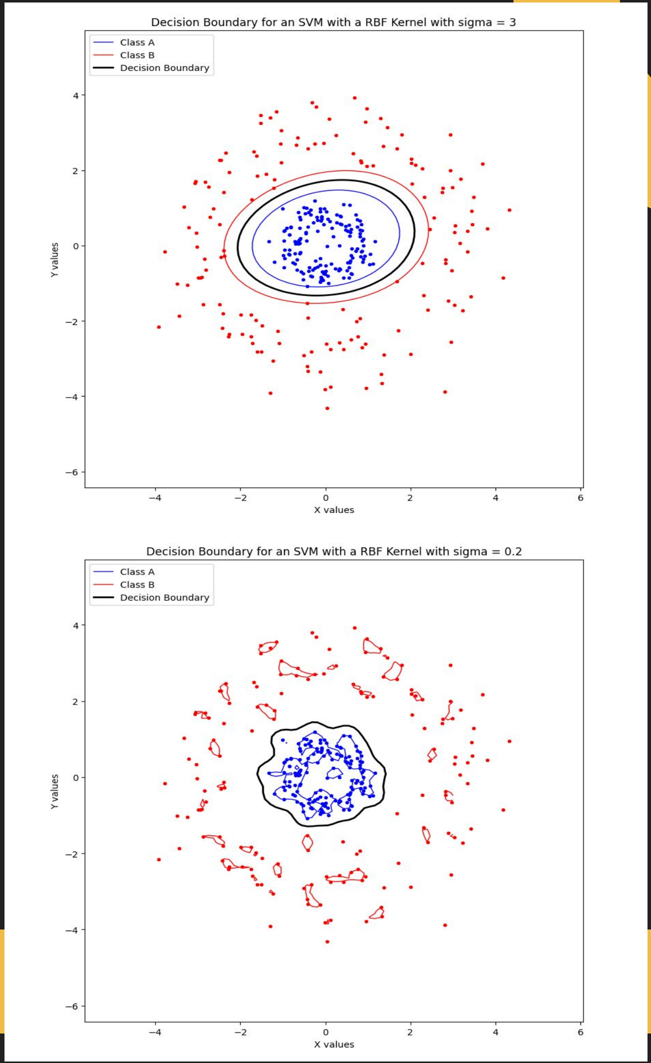
RBF kernels compute the similarity between two points in a high-dimensional feature space.

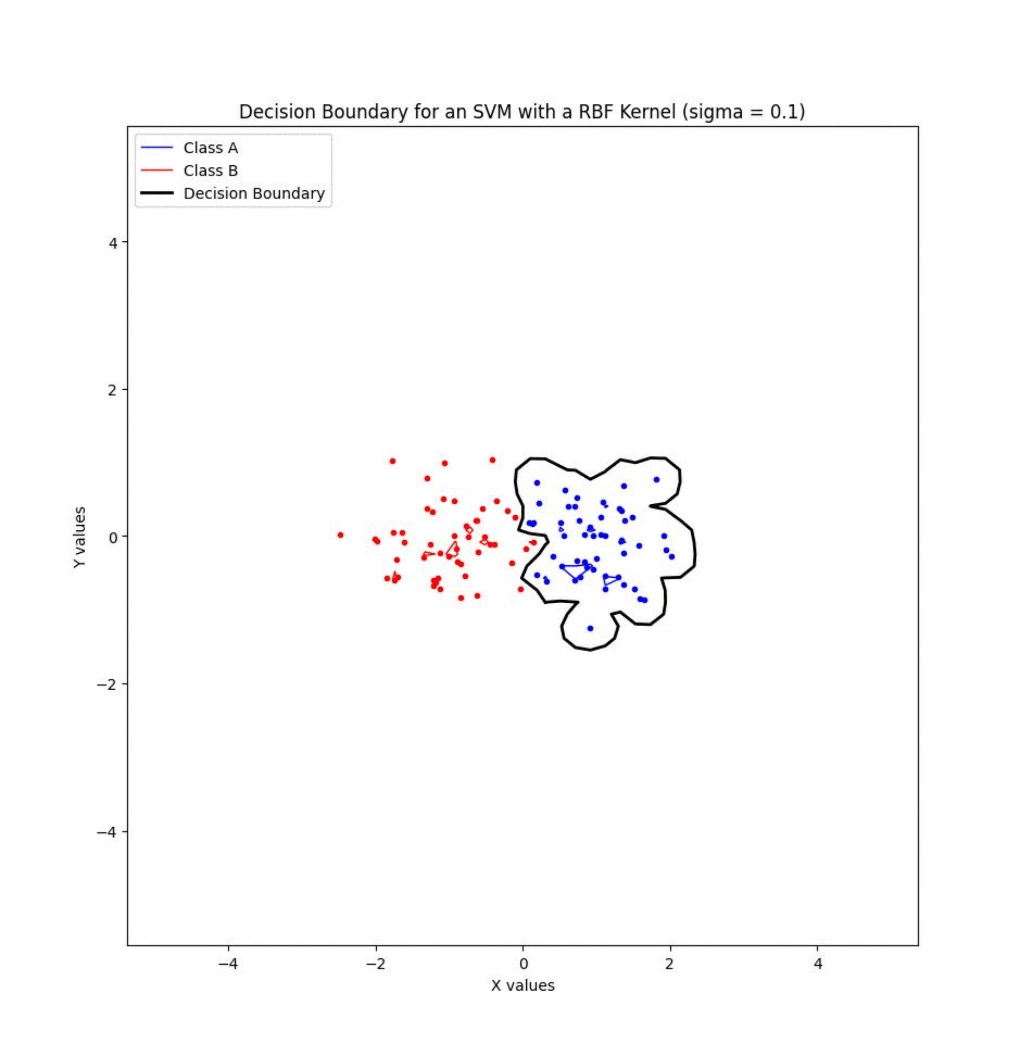
Sigma (σ) is a hyperparameter for the transformation, representing the bandwidth of a Gaussian function.

In terms of bias-variance:

- Increasing σ :
    - Decreases the variance, smoothes the decision boundary, and is less capable of capturing complex data. This also can lead to an increase in bias.

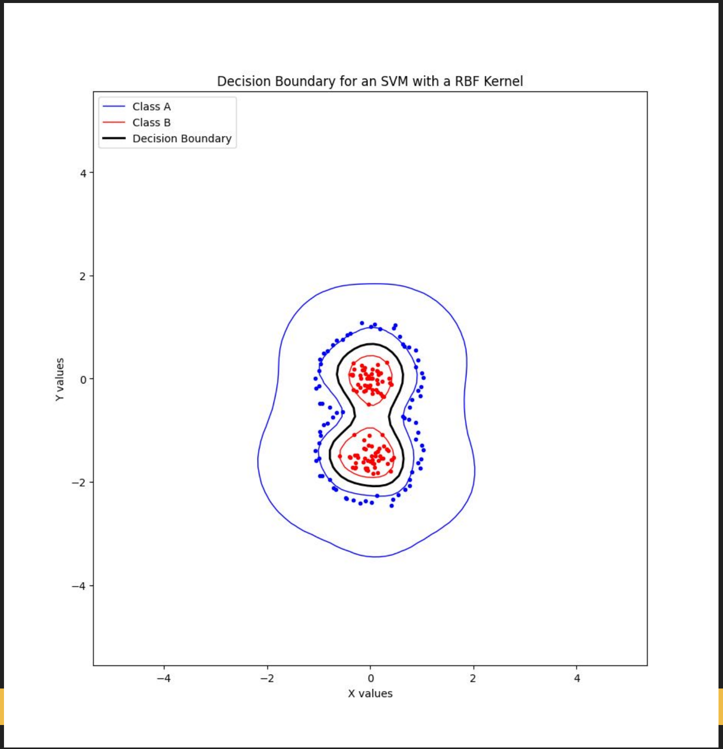
- Decreasing σ :
    - Increases the variance, as the decision boundary has a greater capacity to fit more complex data. Leads to a decrease in bias.



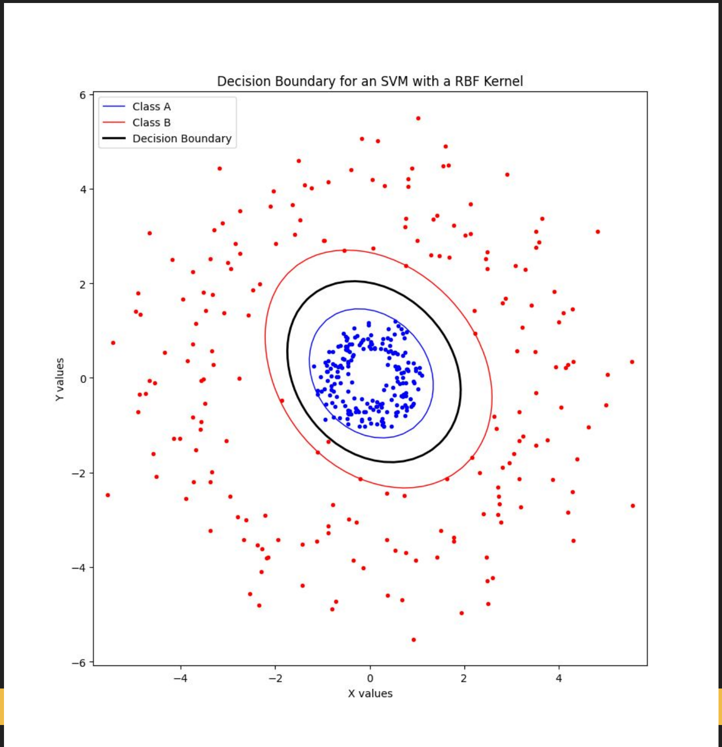

Decision Boundary for an SVM with a RBF Kernel (sigma = 0.1)
Class A
Class B
Decision Boundary
Y values
X values
4
2
0
−2
−4

RBF KERNELS

$$K(X,Y) = \frac{||X - Y||^2}{2\sigma^2}$$



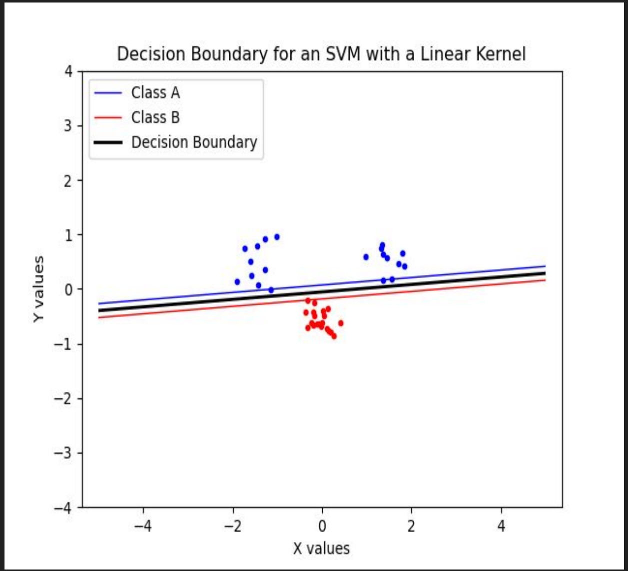
Examples of decision boundaries created with RBF kernels.



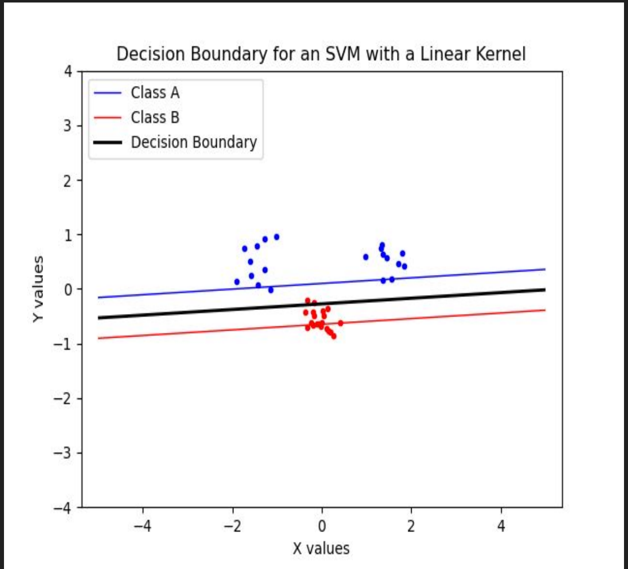
SLACK

- If classes of data are **not** linearly separable, misclassification is unavoidable.
- Slack is parameterised by C, where C controls the trade-off between minimising the sum of the slack variables and maximising the margin.
- This allows the SVM to become tolerant to some amount of misclassification.
- A larger C results in smaller margins and less misclassification.
- A smaller C results in larger margins and more misclassification.

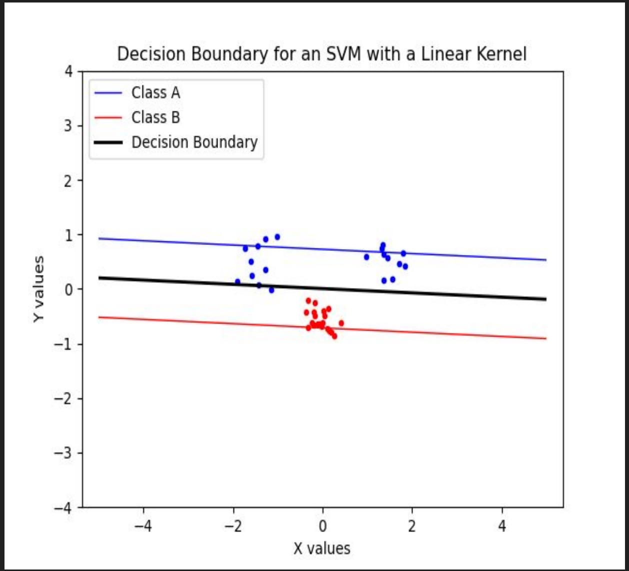
COMPARISON OF DIFFERENT C VALUES



C = 100



C = 1



C = 0.1

SLACK VS COMPLEXITY

- Decreasing the slack parameter (C) can make the model more **flexible** and allows for some misclassification. It also reduces the **potential of overfitting** occurring.

- In general, if training data contains a **lot of noise**, increasing the amount of **slack (decreasing C) is often better**, since it allows the model to be more forgiving of misclassified data points.

- However, if the data has **little noise and high accuracy is extremely important**, using a **more complex model** is often a better choice than introducing slack.

THANK YOU FOR LISTENING!