Machine Learning – Lab 2
SVM

-Shounak Chakraborty

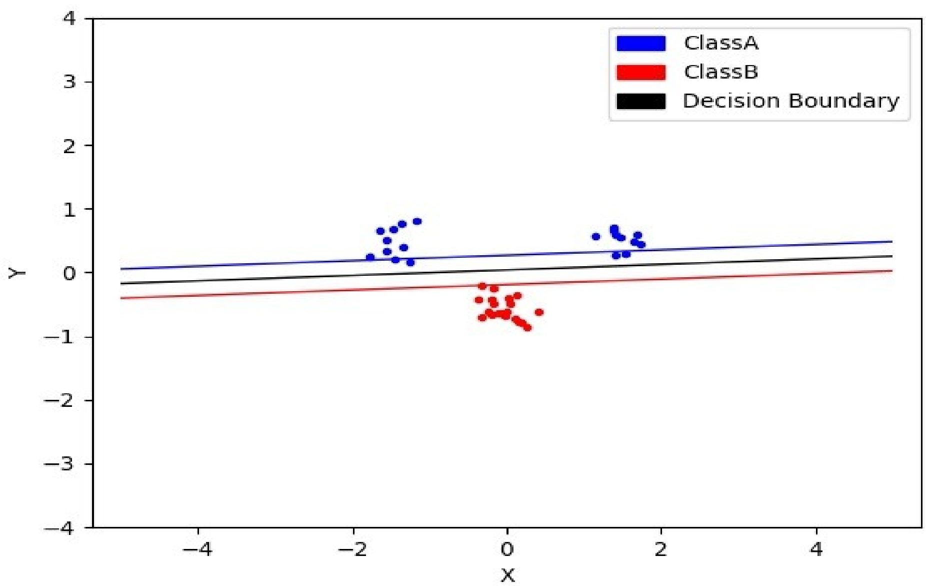
- You need to write code for structuring the data so that the library routine can find the maximal-margin solution, and code for transforming this solution into a classifier which can process new data.

Exploring and Reporting

- **Indicator function** = makes it possible to classify new data points.If the indicator returns a positive value, we say that $\vec{s}$ belongs to class 1, if it gives a negative value, we conclude that the class is −1. All the training data should have indicator values above 1 or below −1, since the interval between −1 and 1 constitutes the margin.
- **Support vectors** = non-zero vectors corresponding to values of alpha, located exactly on the margins, giving support in a mechanical sense.
- **Kernel trick** = allows us to use transformations into very high-dimensional spaces without the penalty of excessive computational costs.
- **Slack variables** = In many cases, especially when the data contain some sort of noise, it is desirable to allow a few datapoints to be miss-classified if it results in a substantially wider margin. This is where the method of slack variables
- **C** = sets the relative importance of avoiding slack versus getting a wider margin

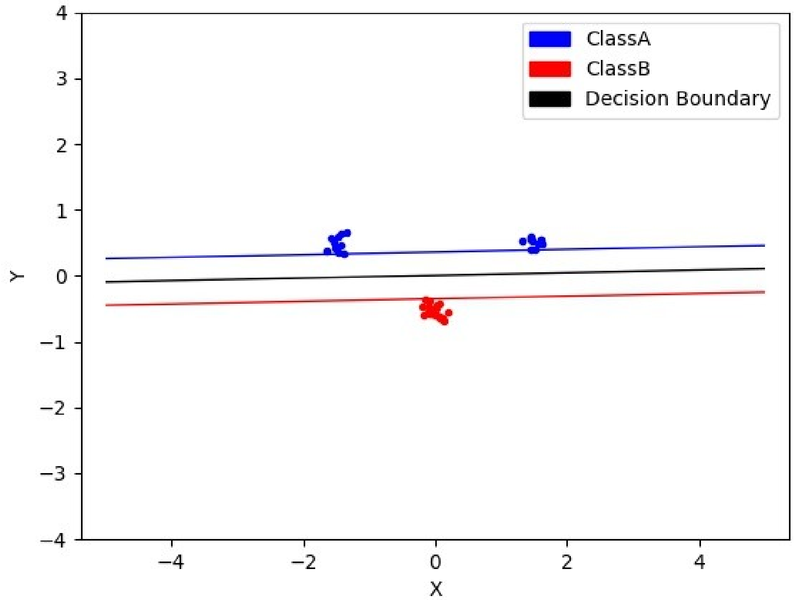
Linear Kernel

C = 10, Standard Deviation of 0.1

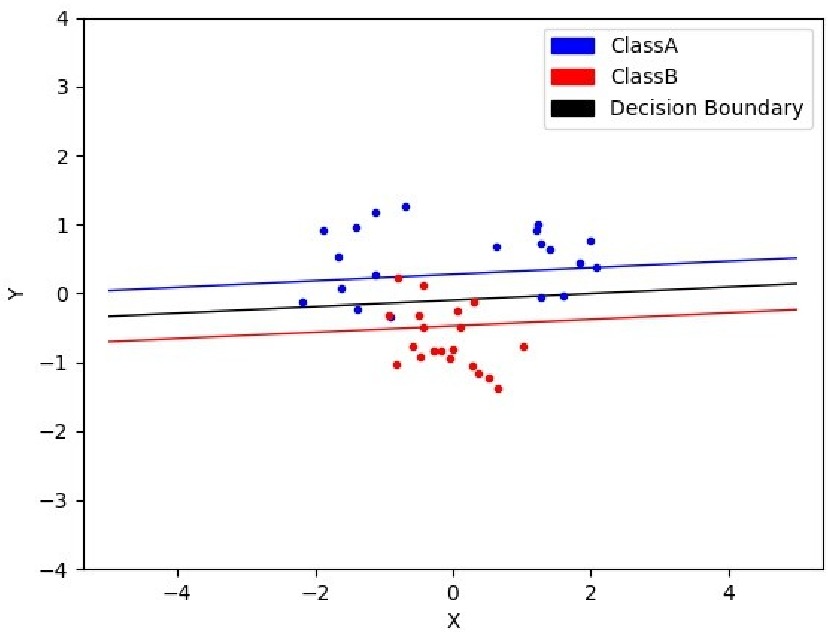


Q.1. Move the clusters around and change their sizes to make it easier or harder for the classifier to find a decent boundary. Pay attention to when the optimizer (minimize function) is not able to find a solution at all.

→ **Standard Deviation of 0.2**
→ **Optimized Successfully**

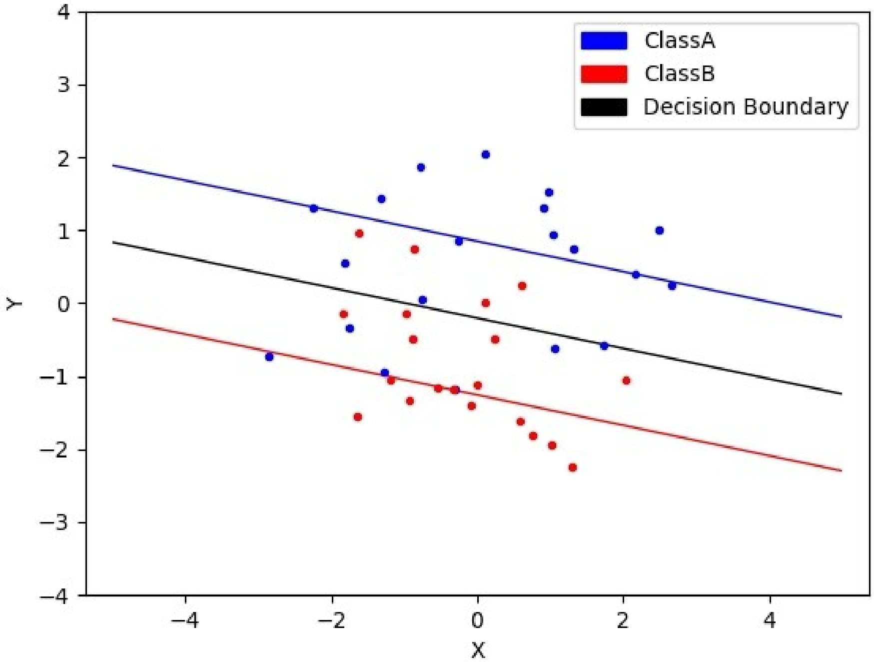


→ **Standard Deviation of 0.5**
→

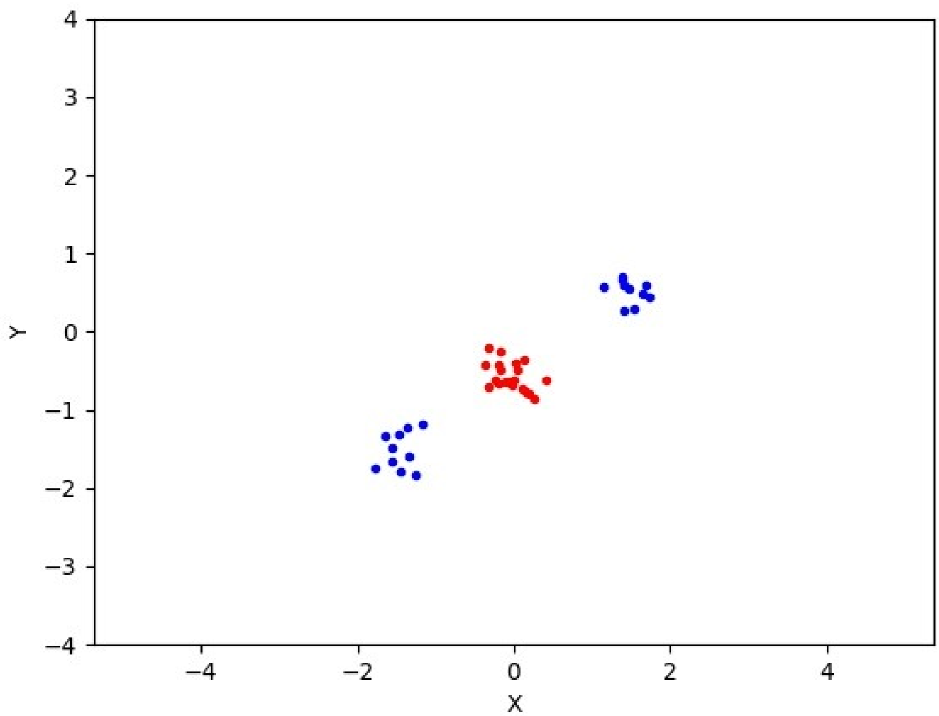


→ **Standard Deviation of 1**

→



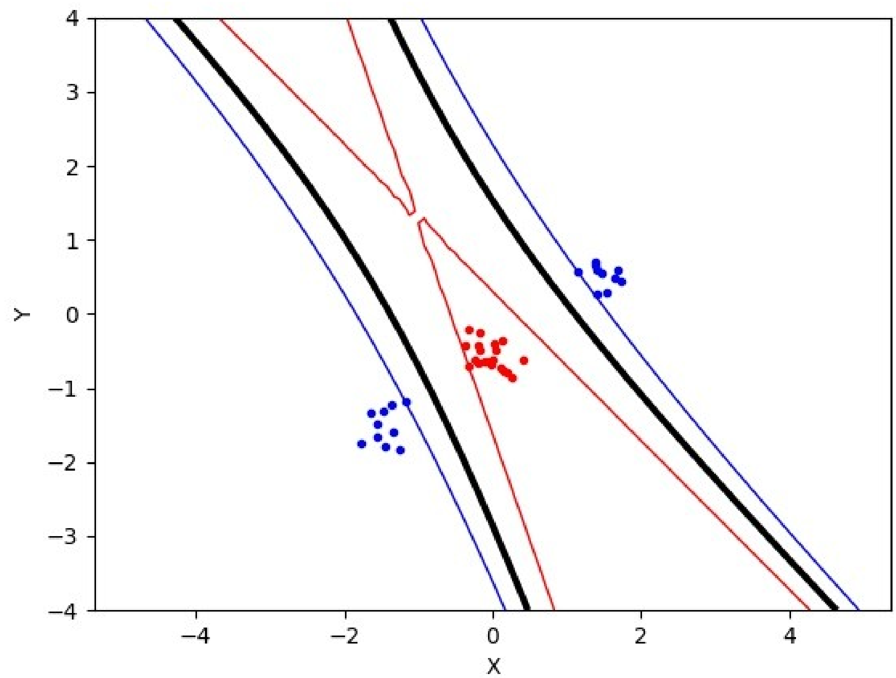
→ **On moving and changing the clusters**



Observations : The Data isn't linearly separable anymore.

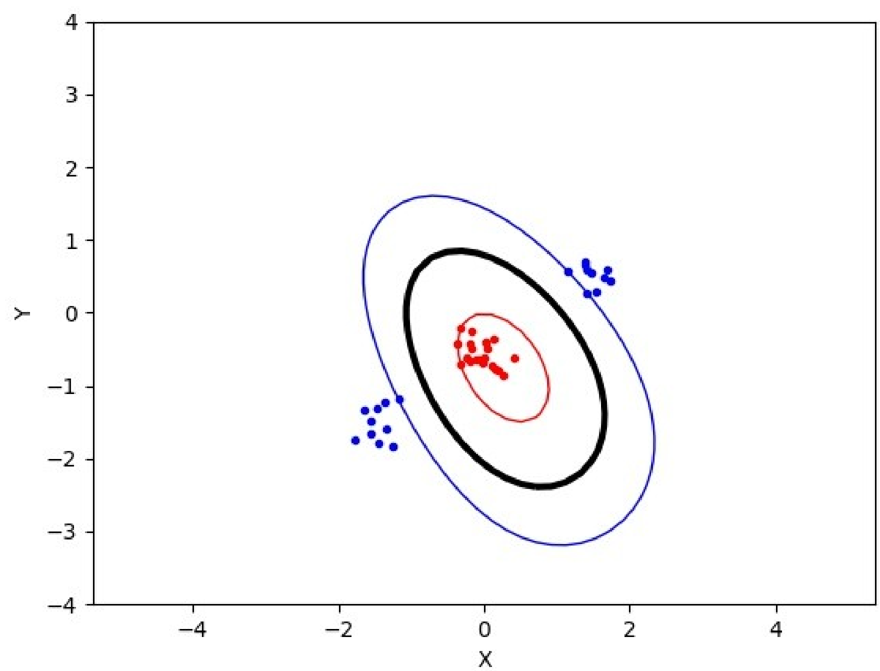
Q.2. Implement the two non-linear kernels. You should be able to classify very hard data sets with these.

→ Using the the above cluster against Polynomial kernel , p=2



→ Using the above cluster woth RBF , WITH SIGMA = 2

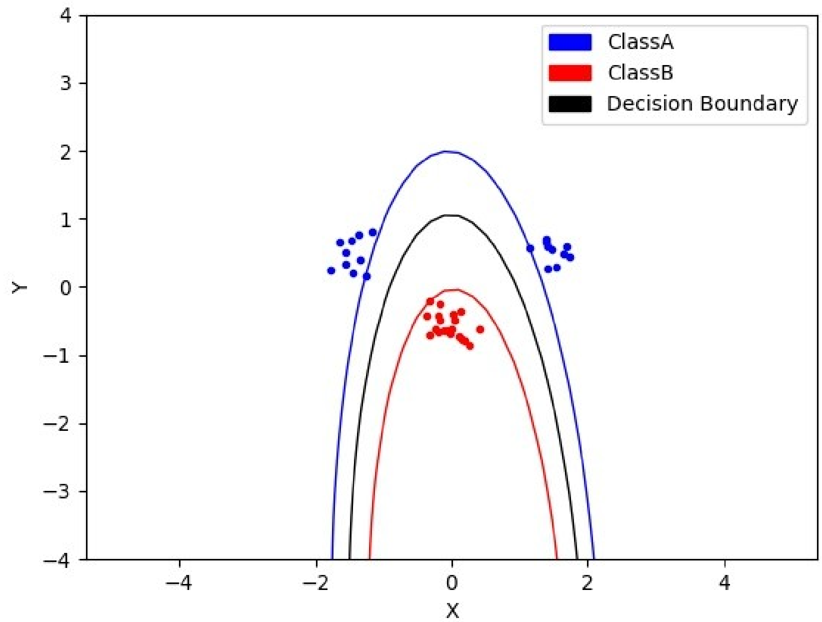
Q.3. The non-linear kernels have



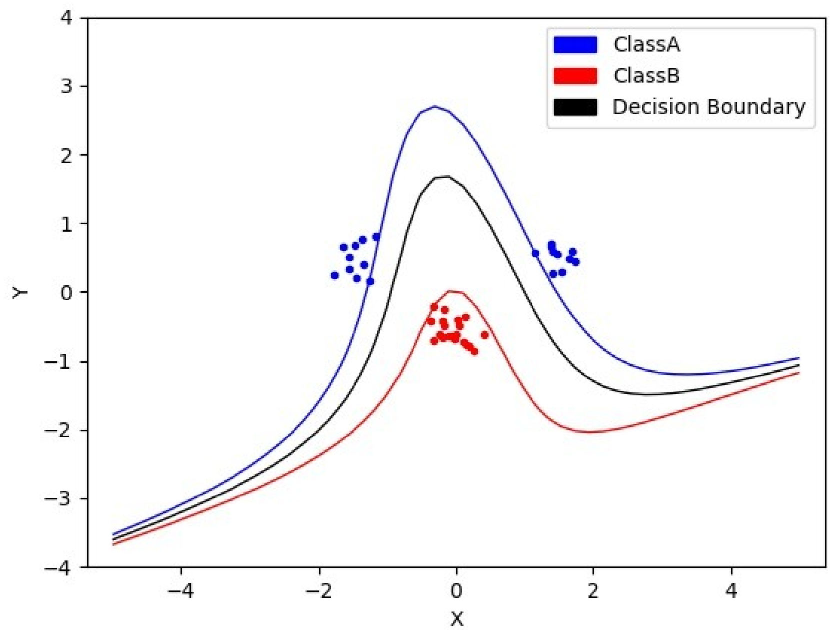
parameters; explore how they influence the decision boundary. Reason about this in terms of the bias- variance trade-off.

Polynomial Kernel

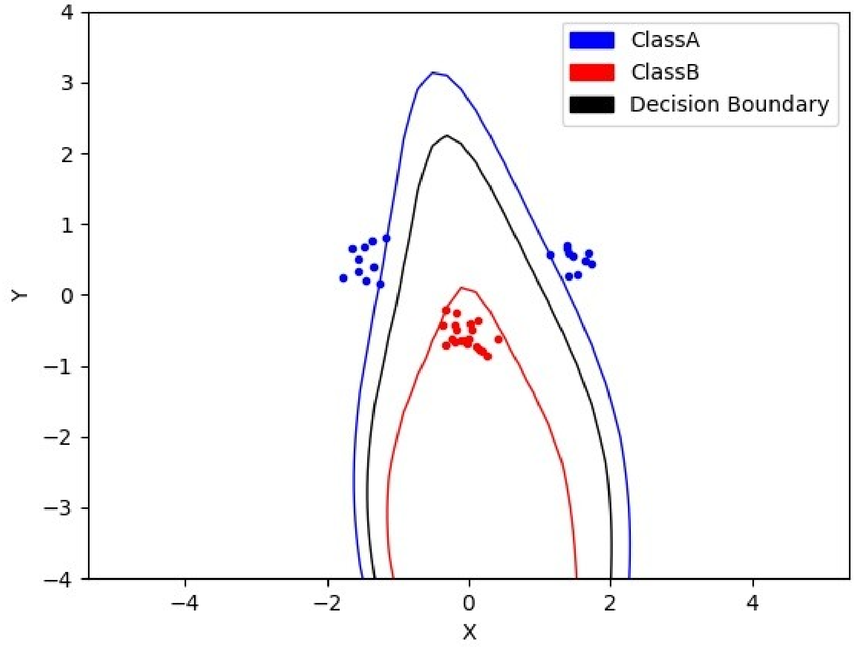
→ **2nd Order Polynomial**



→ **3rd Order Polynomial**



→ **4th Order Polynomial**

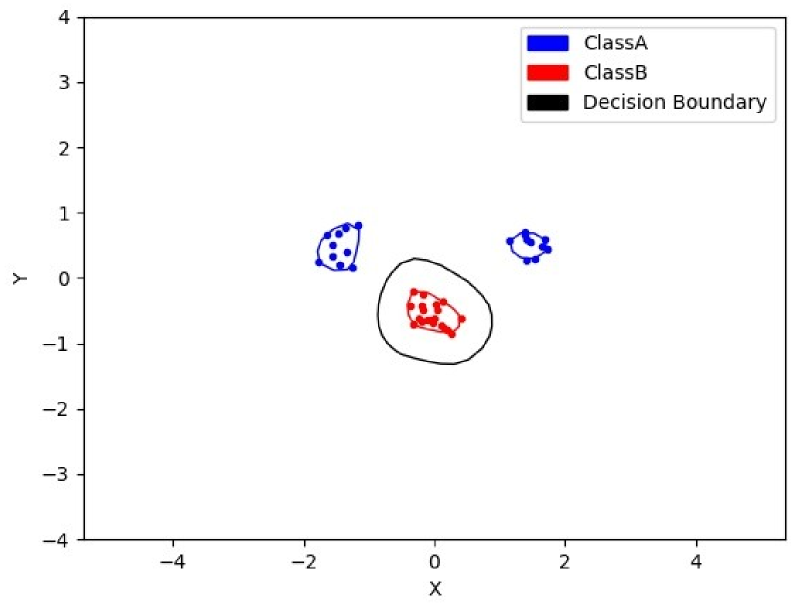


Observations:

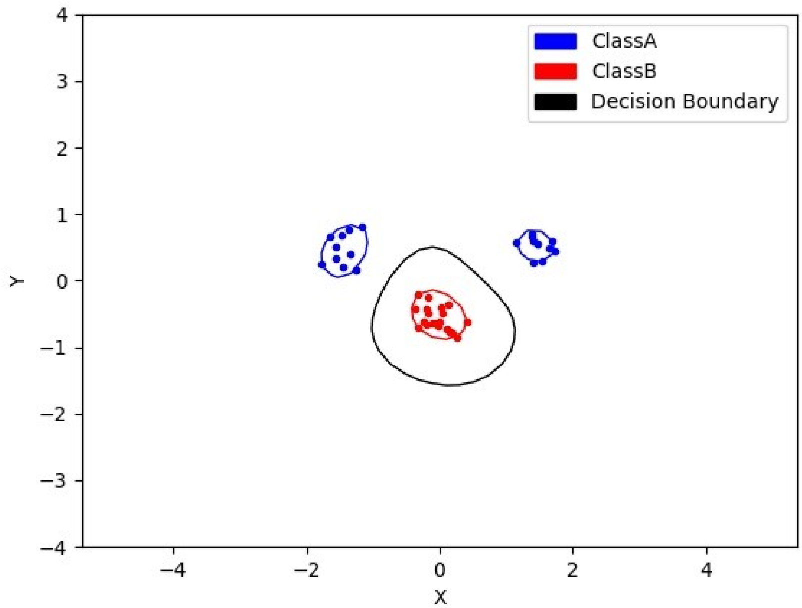
- As p increases we are adding dimensions to the feature space
- The complexity of the decision boundary increases thus increasing the variance and decreasing the bias

Radial Basis Function Kernel

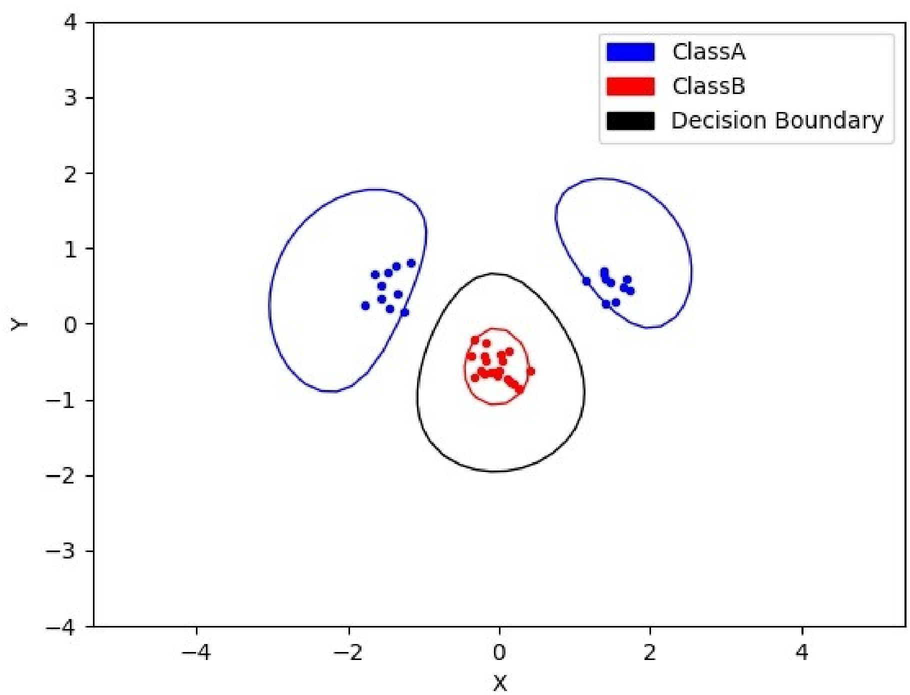
→ **sigma = 0.3**



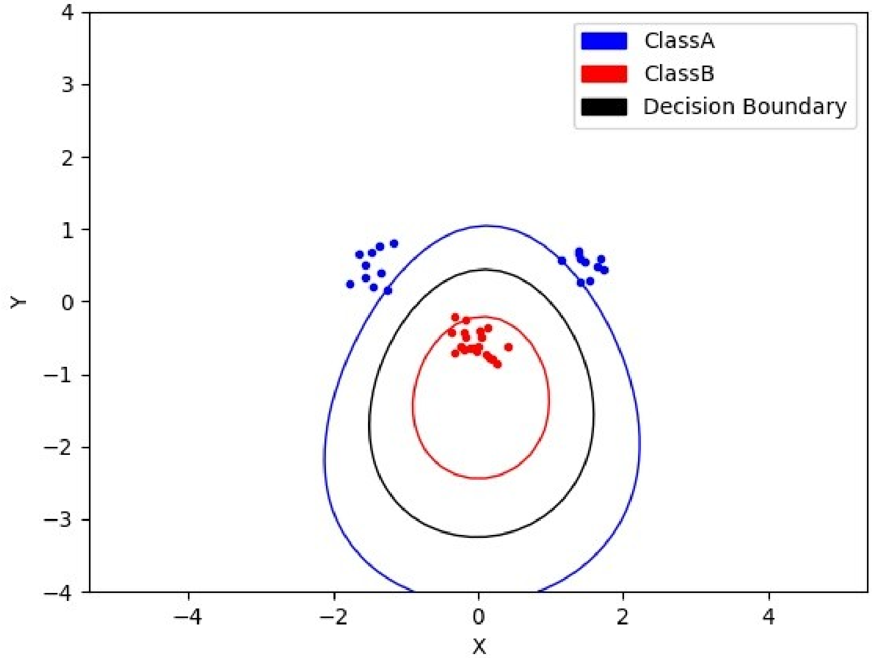
→ **sigma = 0.5**



→ **sigma = 1**



→ **sigma = 2**

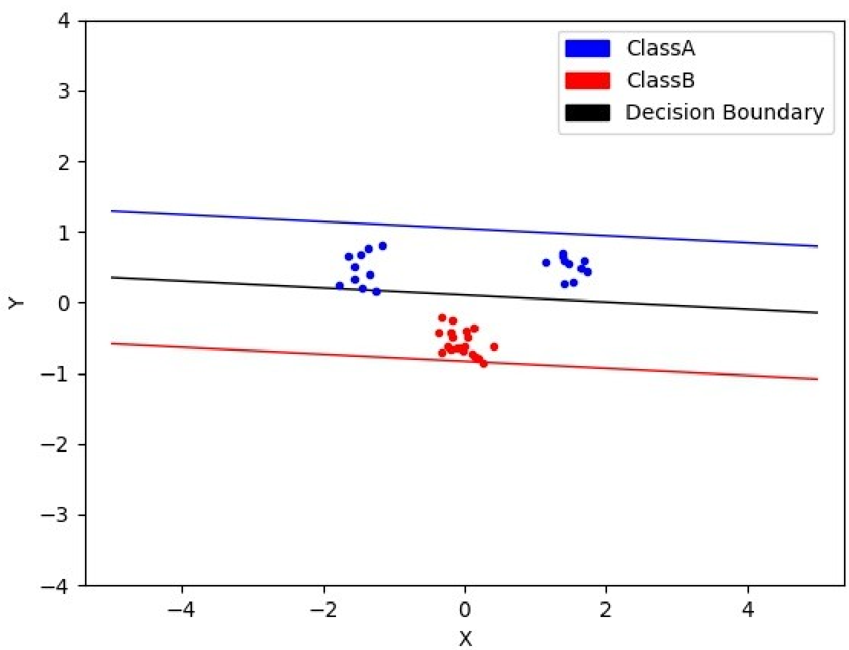


Observations :

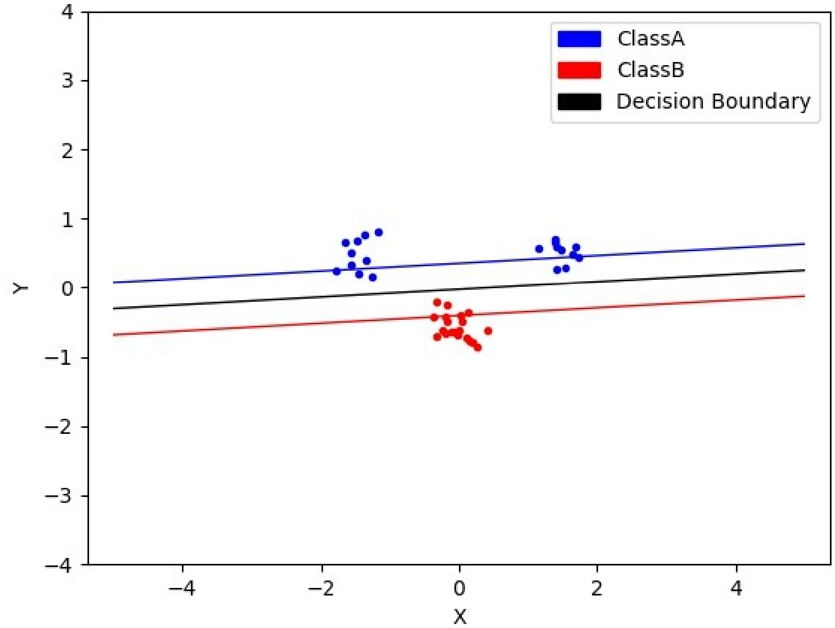
- As sigma increases we are increasing the smoothness of the boundary
- As the complexity decreases so does variance increasing the bias

Q.4 Explore the role of the slack parameter C. What happens for very large/small values?

→ **Linear Kernel**
→ **C= 0.05**

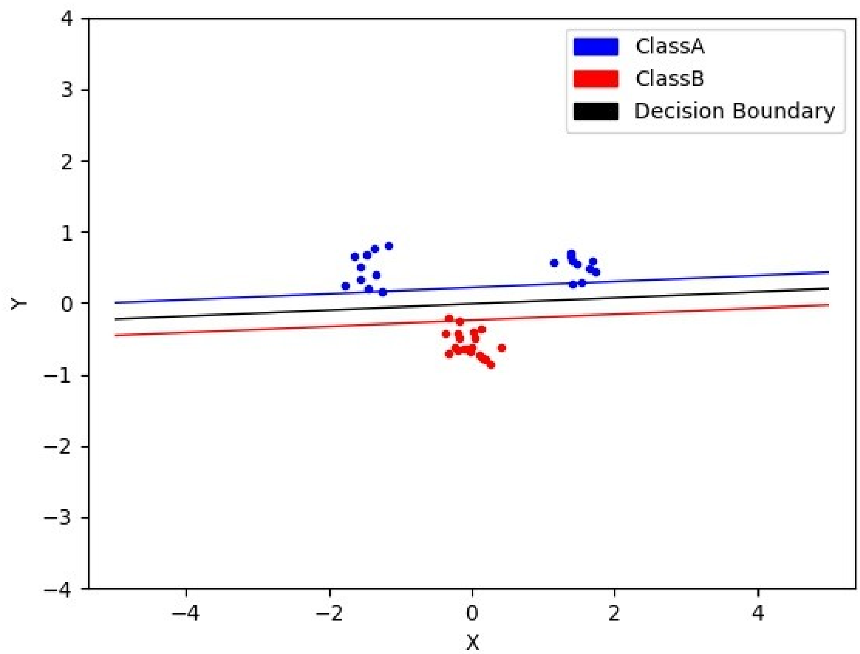


→ **C=1**



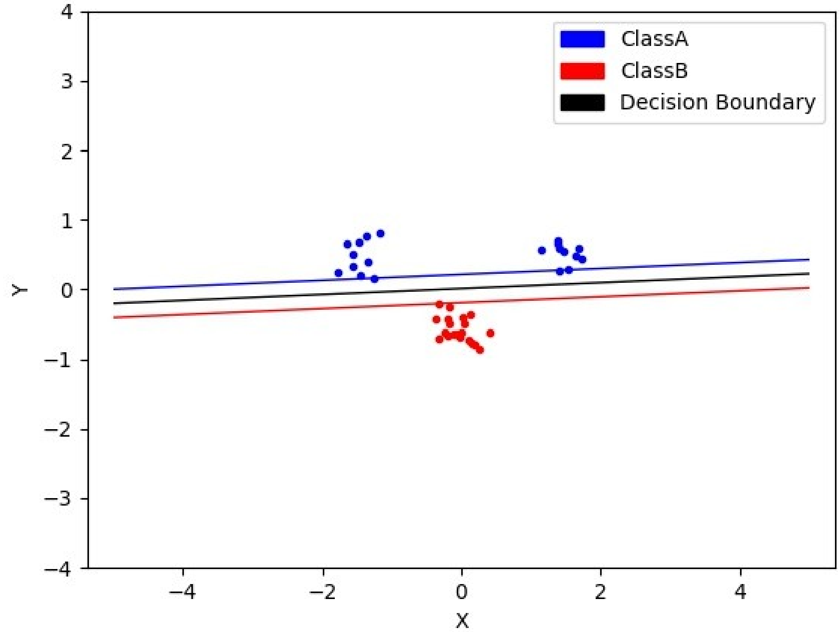
ClassA
ClassB
Decision Boundary
Y
X
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

C =10



ClassA
ClassB
Decision Boundary
Y
X
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

C = 1000



Observations:

-The slack parameter indicates how acceptable it is to have miss classifications. The data points inside the margin are given less importance.
-As we increase C, we are decreasing the tolerance for the error, thus also decreasing the margins.
-As we decrease C, margins become larger and include more data points as we tolerate more error.
-Very high values of C might lead to overfitting, while low values can lead to underfitting even in linearly separable data

Q.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?

- The number of support vectors should not be very high, as that would represent overfitting.
- If the model shows a very high number of support vector, then we should increase the slack in order to reduce the margin or try a simpler model.
- The same rationale applies vice-versa.