DD2421: Machine Learning

Lab 2 : Support Vector Machines

Amna Irshad (amnai@kth.se)
Aravind Nair (aanair@kth.se)

March 24, 2023

KTH Royal Institute of Technology
Stockholm, Sweden

Assignment 1 and 2

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

Changing the values of x and y in numpy.random.randn$(10, 2) * 0.2 + [\mathrm{x}, \mathrm{y}]$ moves the cluster. When the clusters of class A and B are overlapping, the optimizer using linear kernel is not able to find a solution at all. For instance, when both clusters of class A are moved as, class1A $= [1, 0.1]$, class2A $= [-1.5, -0.5]$)) classB $= [0.0, -0.5]$ The optimizer can not find the solution for linear kernel, and gives the error. However, for polynomial kernel, the optimizer gives a distinct boundary, as shown in figure 1,

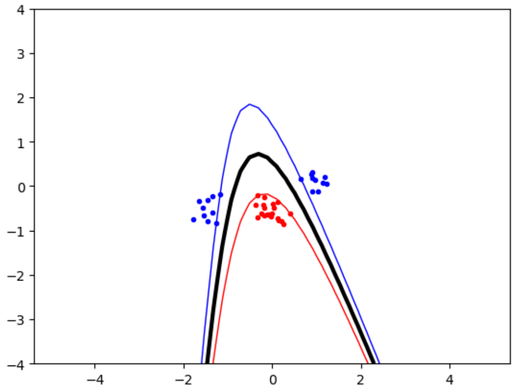


Figure 1: Boundary for polynomial Kernel with p=2

Similar to previous case, if two clusters of class A with standard deviation of 0.4 are placed [1, -0.5] and [-1.5, -0.5] and cluster of class B is at the same place as before, but with standard deviation of 0.3, then the classified shows error, when p is varied from 2 to 5. It only classifies when p=6, as shown in figure 2.

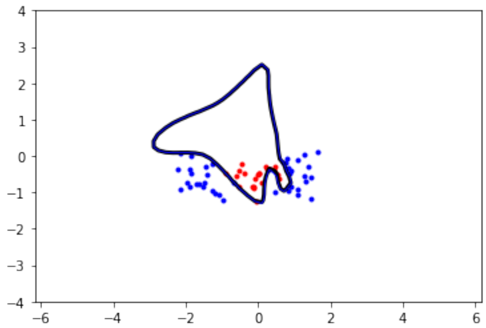


Figure 2: Boundary for polynomial Kernel with p=6

Using the RBF kernel for the clusters at the same positions, the sigma is set to 0.2. For sigma greater than zero, classifier does not give an optimum boundary.

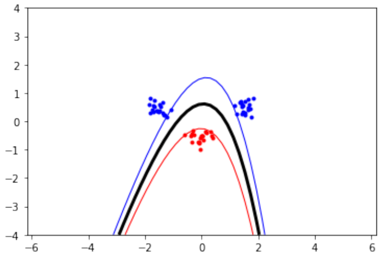
Assignment 3

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

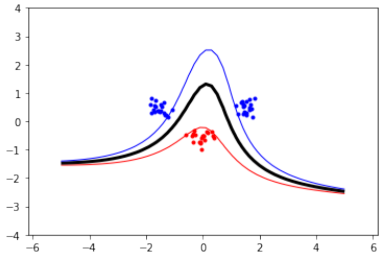
There are two nonlinear kernels discussed specifically Polynomial and RBF.

0.1 Polynomial Kernel

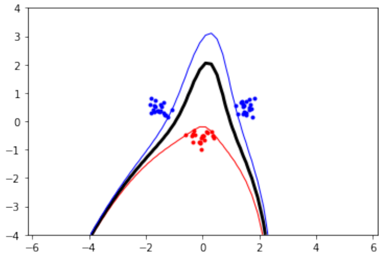
Polynomial kernel has the parameter, commonly denoted by p(a positive integer), controls the degree of the polynomials. $p = 2$ will make quadratic shapes like ellipses, parabolas, hyperbolas. Setting $p = 3$ or higher will result in more complex shapes. For the original data on class A and B, the variation of p is depicted in the figure 3. It is observed, that increasing the value of p, leads to more complex shapes.



(a) $p = 2$



(b) $p = 3$

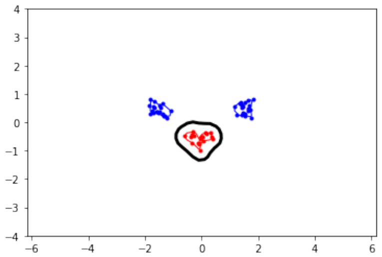


(c) $p = 4$

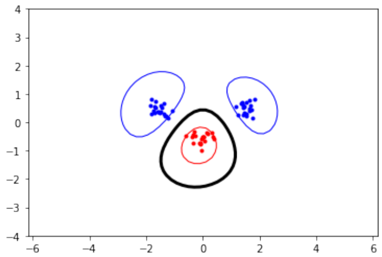
Figure 3: Effect of changing parameter in polynomial kernel

0.2 Radial Basis Function kernel

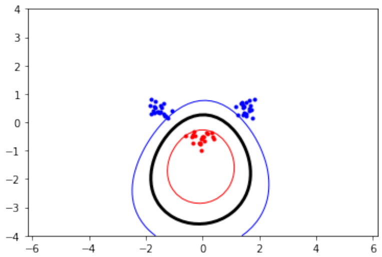
The RBF kernel has a parameter σ which controls the smoothness of the boundary, while RBF uses the explicit euclidean distance between the two data points. The effect of varying value of sigma is more prominent in the figure 4.



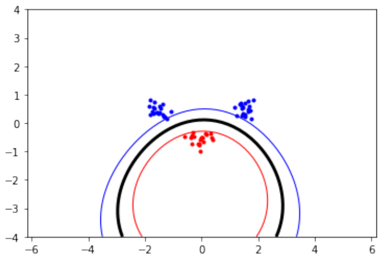
(a) $\sigma = 0.2$



(b) $\sigma = 1$



(c) $\sigma = 2$



(d) $\sigma = 3$

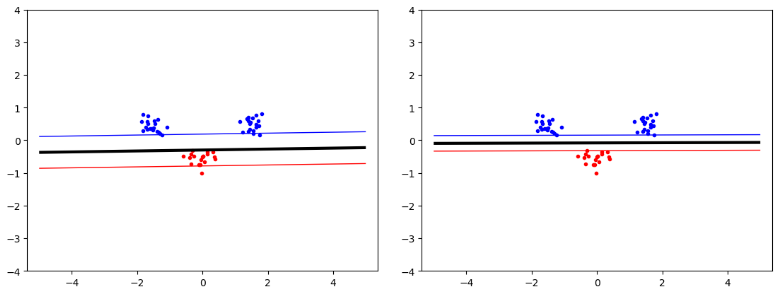
Figure 4: Effect of changing parameter in RBF kernel

It is observed that decreasing sigma leads to more fitted boundary that may lead to over fitting. However, increasing sigma leads to smoother boundary considerations and may lead to better generic results.

1 Assignment 4

Explore the role of the slack parameter C. What happens for very large/small values? The parameter C explores the trade-off of avoiding slack versus getting a wider margin. More slack means that classifier can make more misclassification errors in the training data, which is suitable for a noisy dataset, which leads to a low C value, since individual data points in strange locations should not be taken too seriously.

In cases when classifier was unable to give an optimized boundary condition using linear kernel at high values of C, if value of C is reduced and more slack is allowed, then the boundary condition is obtained at lower value parameters. In linear kernels, changing the value of C predominantly only affects the margins, as depicted in figure 5.



(a) $C = 0.2$ (b) $C = 10000$

Figure 5: Effect of varying parameter C for linear kernel

2 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 easily separable, then a linear model can capture the hidden relation between features and classes more correctly. However if the data is not easily separable, then considering more slack with some missclassification errors allowed, can depict a rather accurate generalization. The bias variance trade-off can be used to explain this relation as well. A complex model or kernel has higher variance and low bias, while more slack represents hogher bias and lower variance.