# Support Vector Machines - Lab 2

Irene Natale

June 4, 2018

## 1 Background

The aim of this lab is to gain knowledge about Support Vector Machines, making use of the dual formulation of the hyperplane theory for classification.
We shall consider a problem with two classes, coded as -1 and +1. Given a new point $x$, the classification of the latter is made thanks to the indicator function: the new point is classified as class -1 in case the value of the indicator function computed at the new point is less than zero, and as class +1 in the opposite case. The indicator function is defined as follows

$$ind(\bar{s}) = \sum_i \alpha_i t_i K(\bar{s}, \bar{x_i}) - b \tag{1}$$

where $\bar{s}$ is the new point that needs to be classified, $t_i$ are the target class of the training data $\bar{x_i}$, b is the offset of the hyperplane and $K(\bar{x}, \bar{y})$ is the kernel function (which can have linear, polynomial or radian form). The $\alpha_i$ are the coefficients that minimize the following optimization problem

$$\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j t_i t_j K(\bar{x_i}, \bar{x_j}) - \sum_i \alpha_i \tag{2}$$

It is also possible to add constraints to the dual formulation of the problem (2), such as

$$0 \leq \alpha_i \leq C \qquad \sum_i \alpha_i t_i = 0 \tag{3}$$

where the parameter C sets the relative importance of avoiding large slack.
The three kernels that will be used in this report are: linear kernel

$$K(x, y) = x^T \cdot y \tag{4}$$

polynomial kernel

$$K(x, y) = (x^T \cdot y + 1)^P \tag{5}$$

and radial kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}} \tag{6}$$

The nature of the used kernel characterizes the "shape" of the final classification, as we will observe in the examples.

In this report we will analyze how linear, polynomial and radial kernels work with three different data sets of increasing difficulty. We will also understand how the parameter C influences the final classification.

## 2 Dataset 1

The first data set that will be subjected to classification is the one shown in Figure 1. The training data points are divided into two classes, one plotted as red (coded as -1) and the other one plotted as blue (coded as +1). In this case the data points are well divided into three distinct groups. The training data shown in Figure 1 are generated from 3 different normal distributions with means and variances $((-1.5, 0.5), 0.2)$, $((1.5, 0.5), 0.2)$ and $((0.0, -0.5), 0.2)$.

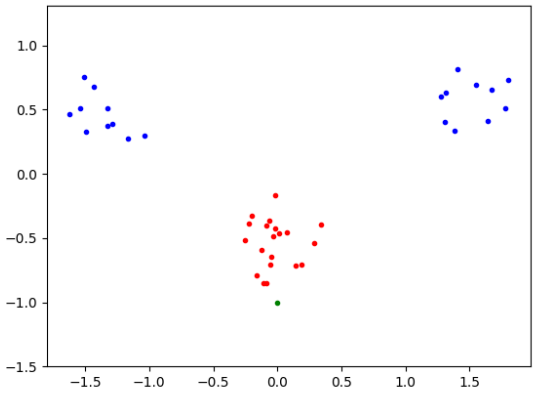



Figure 1: Data1

We first try to classify new data points with a linear kernel. A linear kernel only allows hyperplanes defined by straight lines, and this is what we see in Figure 2. The fact that the linear kernel is not able to summarize the training data points is quite straightforward.
The black lines represent the points of the plane that have indicator function value equal to zero. Points that return value -1 are on the blue line, and points that return value +1 are on the red line. This legend will be used in the whole report.

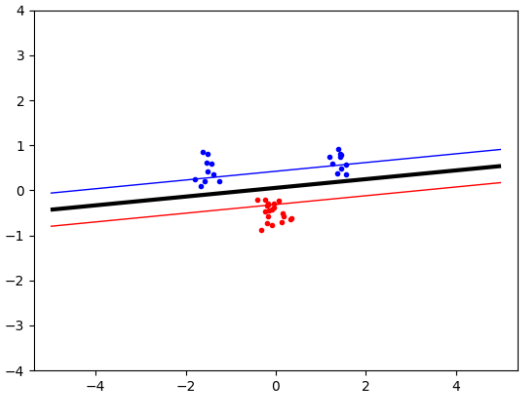
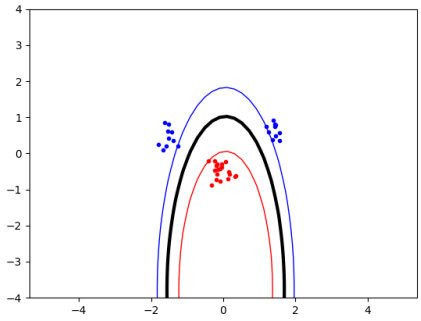
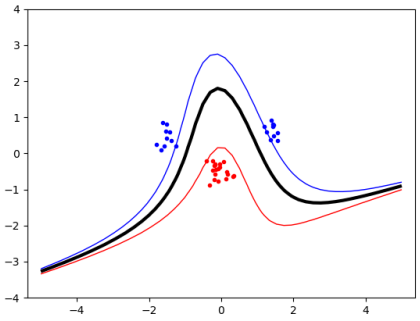



Figure 2: Data1, Linear Kernel, C=0.7

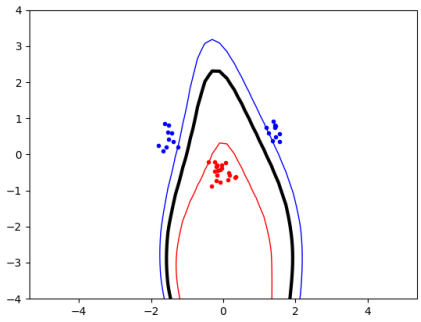
Figure 3 shows how the polynomial kernel behaves. Varying the degree of the polynomial kernel, the model behavior changes. The figure shows the decision hyperplanes for the cases with degree 2,3,4,5. It is possible to notice that the degree of the kernel obviously influences the shape of the hyperplane. The case with $p = 2$ only allows parabolic shapes, but as the polynomial degree increases the sharpness of the curve does the same and the hyperplane assumes a bell shape. With a higher p we are trying to better fit the hyperplane to the training data, but we need to be careful because when p increases, the model gets low bias but high variance, therefore small modifications in the support vectors could significantly change the final solution. At the same time, when p in low, the model has high bias and low variance, making the model not perfectly fit the training data points.



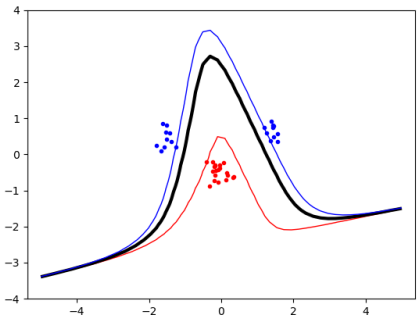
(a) Data1, Polynomial kernel, p=2, C=1



(b) Data1, Polynomial kernel, p=3, C=1



(c) Data1, Polynomial kernel, p=4, C=1



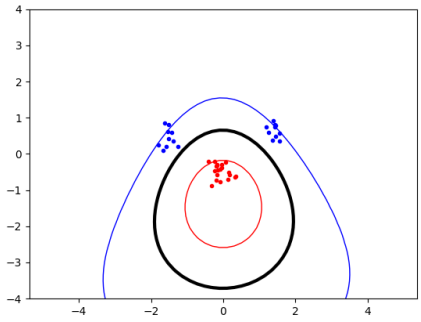
(d) Data1, Polynomial kernel, p=5, C=1

Figure 3: Polynomial kernel

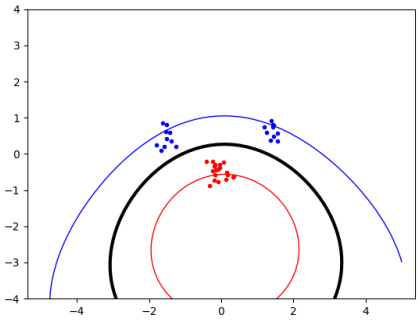
Figure 4 shows how the radial kernel affects the shape of the hyperplane. The radial kernel allows the hyperplane to assume a circular shape. As we showed in (6), the radial kernel has a parameter $\sigma$, which controls the smoothness of the boundary, and so the slope of the hyperplane. When $\sigma$ increases, the smoothness of the boundary increases, and when $\sigma$ decreases the smoothness of the boundary does the same. Therefore, as $\sigma$ increases, the slope of the curve needs to decrease. Note that if $\sigma$ goes to infinity, the radial kernel becomes equivalent to the linear one. When we use a small *sigma*, the model has a low bias and a high variance, while the opposites happens if we choose the use a large $\sigma$.

# 3 Dataset 2

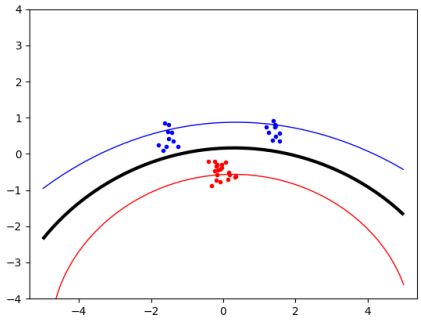
The second dataset on which we work on is the one shown in Figure 5. In this case the data points are generated from three normal distributions with means and variances $((-2, 0), 0.5)$, $((2, 0), 0.5)$ and $((0, 0), 0.5)$. This data set is more difficult to classify than the first data set because the division between different classes is less clean. We are now going to do a study of the parameter C on this data set. We will not use the linear kernel because we saw in the previous example that it does not work well in these kinds of data sets. Therefore we use a radial kernel with fixed $\sigma = 2$ and we change the values of C into 0.2, 0.5, 2 and 4. The plots are shown in Figure 6. The parameter C reflexes the relative importance of avoiding large



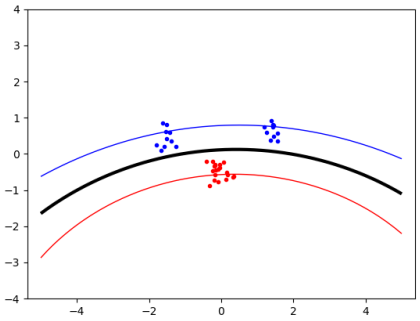
(a) Data1, Radial kernel, sigma=2, C=1



(b) Data1, Radial kernel, sigma=3, C=1



(c) Data1, Radial kernel, sigma=5, C=3



(d) Data1, Radial kernel, sigma=6, C=5

Figure 4: Radial Kernel

slack. This means that for low values of C we will have large slacks with more data points inside, and this is optimal very noisy data sets. Instead, for higher values of C we obtain thiner slacks, with less points inside. This is shown in Figure 6, where it is clear that when C increases, the margins become more narrow.

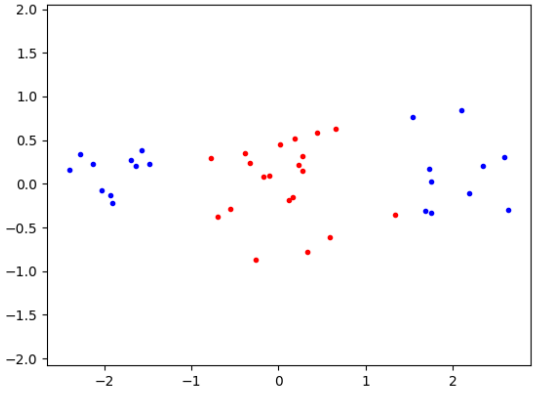
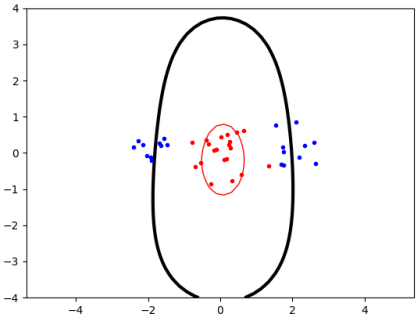
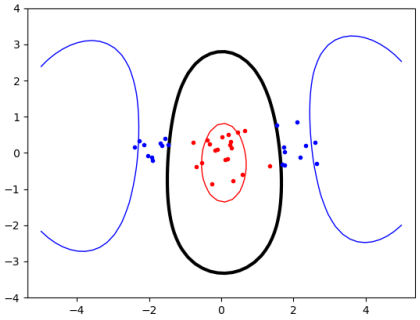


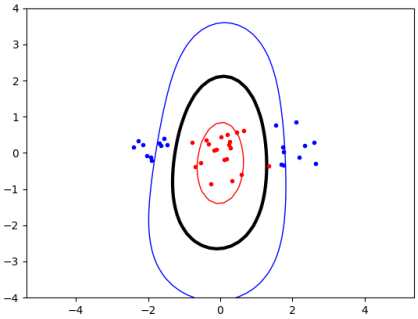
Figure 5: Data2



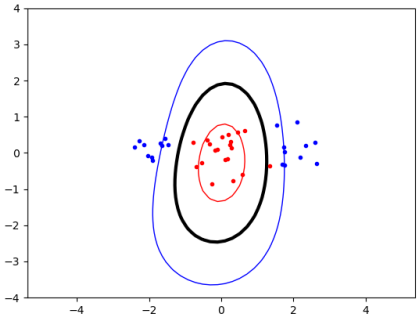
(a) Data2, Radial kernel, sigma=2, C=0.2



(b) Data1, Radial kernel, sigma=2, C=0.5



(c) Data1, Radial kernel, sigma=2, C=2



(d) Data1, Radial kernel, sigma=6, C=4

Figure 6: Radial Kernel

# 4 Data3

In this section the last data set is presented. In this case the data set is very not separable. The data points are generate from three normal distributions with means and variances of $((2,0),0.8)$, $((-2,0),0.8)$ for class A and $((0,0),0.8)$ for class B. The data points are shown in Figure 7.

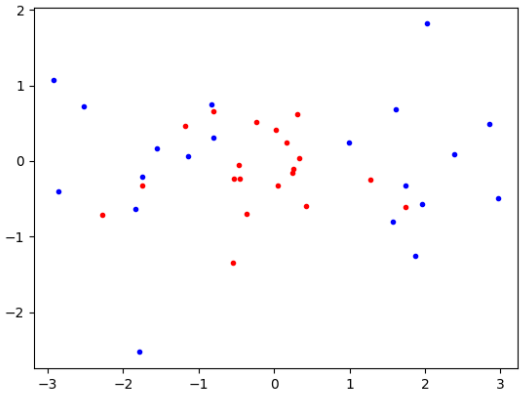
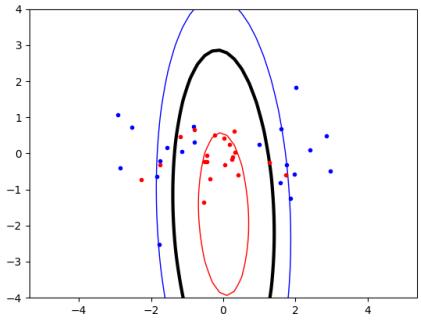


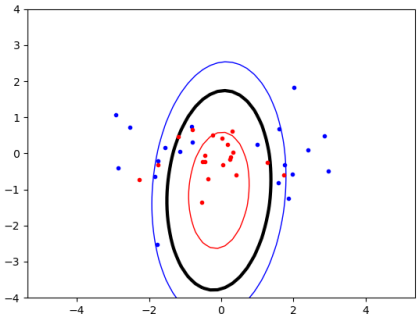

Figure 7: Data3

With this data set we will try to answer to the following question: When should we opt for more slack rather than going for a more complex model (kernel) and vice versa? Figure 8 shows the resulting hyperplane from polynomial kernels of degree 2 and 4, with $C = 0.5$ and $C = 2$ for both cases. In Figure 9 we can also see what we get if we solve the classification problem with a radial kernel with fixed $\sigma = 0.7$ and values of $C = 0.5$ and 8.

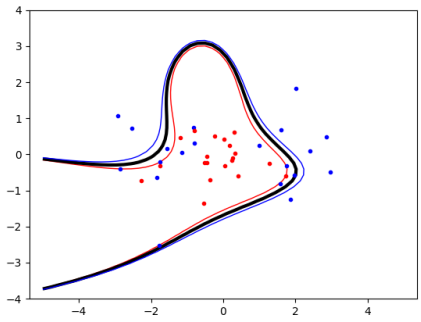
Figure 8.a represents the starting situation, in which we use a polynomial kernel with low C (so large margins), but it does not classify well the data set. If we increase C the model still does not work well enough. It is good in this case choosing to increase the degree of the polynomial kernel, as it is shown in figures 8.c and 8.d. In fact in these cases the model fit better the training data. We still need to pay attention to not use a model that is too complex, for example a radial kernel with a high C (Figure 9.b). In this case the model fits too well the training data and it will not work on the test data set. We will obtain an overfitting situation. In conclusion, in these case I would suggest a wider cross-validation study, in order to find the good balance between model complexity, which, if high, leads to a good fitting, and test data sitting.



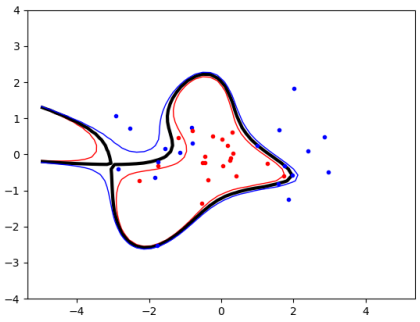
(a) Data3, Polynomial kernel, degree=2, C=0.5



(b) Data3, Polynomial kernel, degree=2, C=2

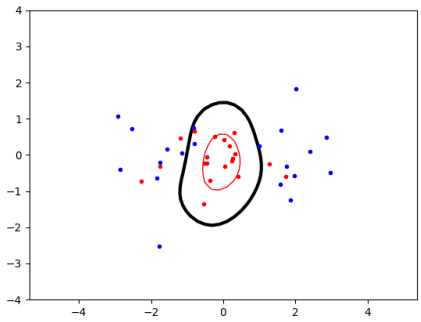


(c) Data3, Polynomial kernel, degree=4, C=0.5

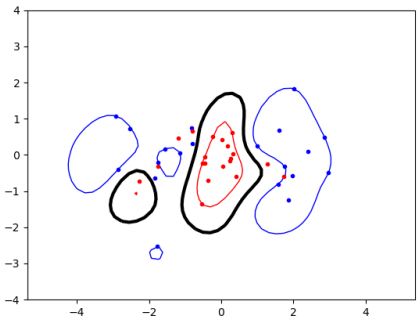


(d) Data3, Polynomial kernel, degree=4, C=2

Figure 8: Polynomial Kernel



(a) Data3, Radial kernel, sigma = 0.7, C=0.5



(b) Data3, Radial kernel, sigma = 0.7, C=8

Figure 9: Polynomial Kernel