Instituto Superior Técnico

Departamento de Engenharia Electrotécnica e de Computadores

2018/2019 – 1st Semester

Machine Learning

5th Lab Assignment

Shift: Wednesday, 14h00

Number: 84037 Name: Eduardo Alexandre Silva da Costa

Number: 84038 Name: Eduardo Miguel Ferreira Cabral de Melo

2. Two simple examples

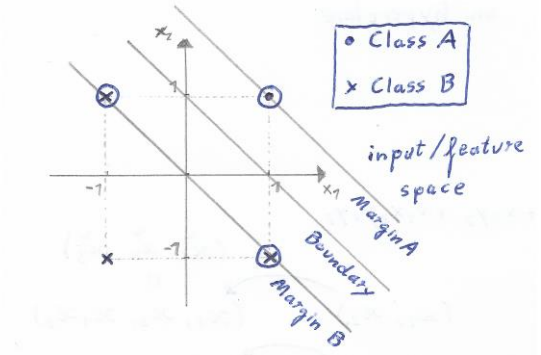
2.1 AND Function

In this case, the feature space will be the input space itself, since it's a separable case. Therefore, the vectors with and without tilde are equal.

x_1	x_2	d_{AND}
-1	-1	-1
-1	1	-1
1	-1	-1
1	1	1

Class A → $d_{AND} = 1$

Class B → $d_{AND} = -1$



Support vectors: S.V.(1,1) ; S.V.(1,-1) ; S.V.(-1,1).

In the graphic above, it's observable the maximum-margin separating straight line, the support vectors (circled) and the margin boundaries. It was used one point of each boundary in order to compute this last, as demonstrated below.

Boundary $x_2 = -x_1 + C$

(1,0) is on boundary

$\Rightarrow 0 = -1 + C \Rightarrow C = 1$

$\boxed{x_2 = -x_1 + 1}$

Margin A $x_2 = -x_1 + C_A$

S.V (1,1) is on margin A

$\Rightarrow 1 = -1 + C_A \Rightarrow C_A = 2$

$\boxed{x_2 = -x_1 + 2}$

Margin B $x_2 = -x_1 + C_B$

S.V(1,-1) is on margin B

$\Rightarrow -1 = -1 + C_B \Rightarrow C_B = 0$

$\boxed{x_2 = -x_1}$

If $\tilde{w} \cdot \tilde{x} + b = 0$ defines the boundary of the maximum-margin, then $(\tilde{w} \cdot \tilde{x}^S + b)d^S = C$, where *C* is a constant.

Also, as stated before, $\tilde{w} \cdot \tilde{x} + b = 0 <=> w \cdot x + b = 0$. Then, considering that the boundary equation is equal to $w \cdot x + b = 0$, it is possible to compute the vector *w* and the bias *b*, as demonstrated below.

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\quad w = \begin{bmatrix} \alpha & \beta \end{bmatrix}$

$\begin{cases} x_2 = -x_1 + 1 \\ w \cdot x + b = 0 \end{cases} \Rightarrow \begin{cases} x_1 + x_2 - 1 = 0 \\ \alpha x_1 + \beta x_2 + b = 0 \end{cases} \Rightarrow \begin{cases} b = -1 \\ \alpha = 1 \\ \beta = 1 \end{cases} \Rightarrow w = \begin{bmatrix} 1 & 1 \end{bmatrix}$ $\quad b = -1$

2.2 XOR Function

x_1	x_2	d_{XOR}
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

Class A → $d_{XOR} = 1$

Class B → $d_{XOR} = -1$

Since our input space isn't linear, we cannot use a linear separation.

This problem is solved by mapping the data from the input space into a higher dimension feature space, so that the different classes of data can be separated by an hyperplane.

$$\tilde{x} = \phi(x) = (x_1, x_2, x_1 x_2)^T \qquad K(x,y) = \phi(x)\,\phi(y)$$

$$K(x,y) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_1 x_2 y_1 y_2$$

2.3

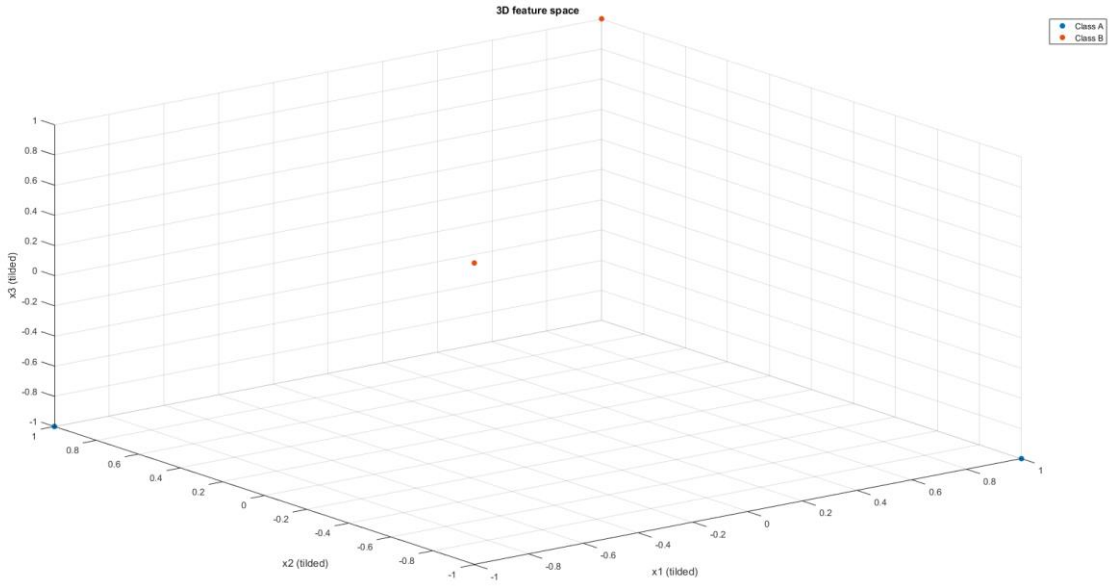


Figure 1

$$(x_1, x_2) \rightarrow (x_1, x_2, x_1 x_2) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$$

$$(-1,-1) \rightarrow (-1,-1,1)$$

$$(-1,1) \rightarrow (-1,1,-1)$$

$$(1,-1) \rightarrow (1,-1,-1)$$

$$(1,1) \rightarrow (1,1,1)$$

In figure 1 we can visualize the points in the feature space. It's observable that class A points have $\tilde{x}_3 = 1$ and that class B points have $\tilde{x}_3 = -1$. Since $x_1 = \tilde{x}_1$ and $x_2 = \tilde{x}_2$, we can use as a boundary the hyperplan $\tilde{x}_3 = 0$ with S.V.(-1,-1,1), S.V.(-1,1,-1), S.V.(1,-1,-1) and S.V.(1,1,1) as support vectors.

Since $\tilde{w} \cdot \tilde{x} + b = 0 \Rightarrow (\tilde{w} \cdot \tilde{x}^S + b)d^S = C$, and considering that the boundary equation ($\tilde{x}_3 = 0$) is equal to $\tilde{w} \cdot \tilde{x} + b = 0$,

$$\tilde{x} = \begin{bmatrix} \tilde{x_1} \\ \tilde{x_2} \\ \tilde{x_3} \end{bmatrix} \qquad \begin{cases} \tilde{x_3} = 0 \\ \tilde{w} \cdot \tilde{x} + b = 0 \end{cases} \Leftrightarrow \begin{cases} \tilde{x_3} = 0 \\ \alpha \tilde{x_1} + \beta \tilde{x_2} + \gamma \tilde{x_3} + b = 0 \end{cases} \Rightarrow \begin{cases} \alpha = 0 \\ \beta = 0 \\ \gamma = 1 \\ b = 0 \end{cases} \qquad \tilde{w} = [0 \ 0 \ 1] \quad b = 0$$

$$\tilde{w} = [\alpha \ \beta \ \gamma]$$

2.4

With the boundary equation and one point from each margin, we can compute both margin boundaries. However, to express them in the two-dimensional input space, we need to change our variables from (,,) to (,,). In this case, since in all equation we only have the component, it's pretty simple. The input space margins are highlighted by rectangles below.

Boundary $\tilde{x_3} = 0 \rightarrow \boxed{x_1 x_2 = 0}$

Margin A $\tilde{x_3} = C_A$ S.V(-1,1,-1) is on margin A $\Rightarrow C_A = -1$

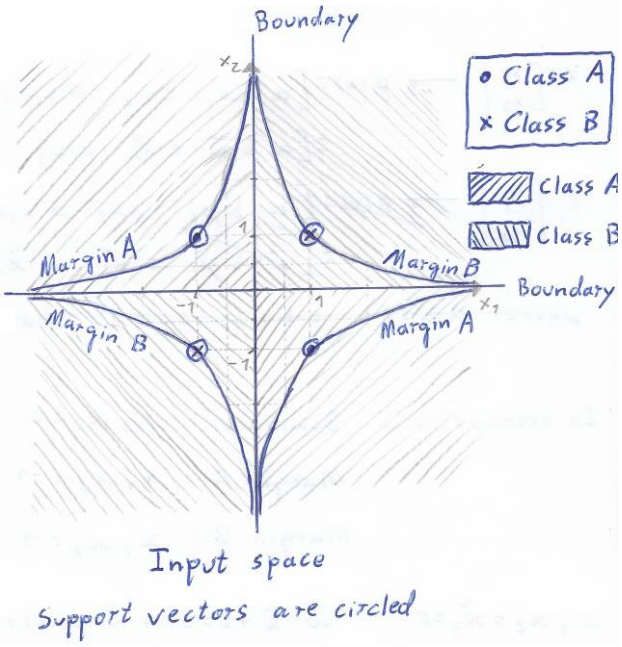
$\tilde{x_3} = -1 \rightarrow \boxed{x_1 x_2 = -1}$

Margin B $\tilde{x_3} = C_B$ S.V(1,1,1) is on margin B $\Rightarrow C_B = 1$

$\tilde{x_3} = 1 \rightarrow \boxed{x_1 x_2 = 1}$

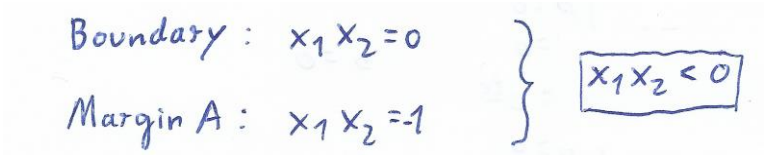
With both margins calculated, we can sketch the input patterns and the margins. We can also define the regions of the input space that the classifier will define as 1 (class A) or -1 (class B).

Note: It may not be very observable, but the maximum-margin boundary consists of both axis x_1 and x_2.



2.5

For a classifier to produce an output of 1, it must belong to class A. With that said,



This condition is also easily observable in the sketch obtained in point 2.4.

3. Classification using SVMs

3.1

Kernel: $K(x, y) = (x \cdot y + a)^p - a^p$, where $a \in \mathbb{R}^+$ and $p \in \mathbb{N}$.

Considering $x, y \in \mathbb{R}^2$, $a = 1$ and $p = 2$

$$K(x,y) = (x \cdot y + 1)^2 = (x_1 y_1 + x_2 y_2 + 1)^2 =$$
$$= x_1^2 y_1^2 + 2(x_1 y_1)(x_2 y_2 + 1) + (x_2 y_2 + 1)^2 =$$
$$= x_1^2 y_1^2 + 2 \cdot x_1 y_1 \cdot x_2 y_2 + 2 x_1 y_1 + x_2^2 y_2^2 + 2 \cdot x_2 y_2 + 1 = \phi(x) \cdot \phi(y)$$

6 dimensions

Dimensionality of the feature space: $\mathbb{R}^6$

Mapping:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longrightarrow \phi(x) = [a \quad b x_1 \quad c x_2 \quad d x_1 x_2 \quad e x_1^2 \quad f x_2^2]^T =$$
$$= [a \quad b \tilde{x}_1 \quad c \tilde{x}_2 \quad d \tilde{x}_3 \quad e \tilde{x}_4 \quad f \tilde{x}_5]^T$$

3.2

In exercise 2:

Boundary	Margin A	Margin B
$x_1 x_2 = 0$	$x_1 x_2 = -1$	$x_1 x_2 = 1$

As stated in the mapping, $x_1 \cdot x_2$ is equivalent to the variable $\tilde{x}_3$. Therefore, our boundary is given by $\tilde{x}_3 = 0$ and our margins will keep the same than in exercise 2, for the same patterns.

Considering the boundary equation $\tilde{x}_3 = 0$ equal to $\tilde{w} \cdot \tilde{x} + b = 0$, where $\tilde{x} = \phi(x)$ we can compute the vector $\tilde{w}$, as demonstrated below.

$$\tilde{w} \cdot \tilde{x} + b = 0 \Leftrightarrow \tilde{w} \cdot \phi(x) + b = 0 \Leftrightarrow$$
$$\Leftrightarrow [\alpha \quad \beta \quad \gamma \quad \delta \quad \varepsilon \quad \zeta] \cdot [a \quad b \tilde{x}_1 \quad c \tilde{x}_2 \quad d \tilde{x}_3 \quad e \tilde{x}_4 \quad f \tilde{x}_5]^T + b_1 = 0 \Leftrightarrow$$
$$\Leftrightarrow \alpha \cdot a + \beta b \tilde{x}_1 + \gamma c \tilde{x}_2 + \delta d \tilde{x}_3 + \varepsilon e \tilde{x}_4 + \zeta f \tilde{x}_5 + b_1 = 0$$

$$x_1 \cdot x_2 = \tilde{x}_3 = 0 \Rightarrow \begin{cases} \alpha = 0 \\ \beta = 0 \\ \gamma = 0 \\ \delta = c^{te} \\ \varepsilon = 0 \\ \zeta = 0 \\ b_1 = 0 \end{cases}$$

$$\tilde{w} = [0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0]$$
$$b_1 = 0$$

Considering δ = 1, for example.

4. Experiments

4.1 Polynomial kernel

Table 1

p	Classification error (%)	Number of support vectors
1	45	100
2	35	100
3	35	99
4	21	78
5	16	91
6	0	35
7	0	82
11	2	99

It's observable in table 1 that with the initial increase of the parameter p, the percentage of the classification error diminishes, such as the number of support vectors. This occurs until $p = 6$. For higher values of p, the number of support vectors increase, and with $p = 11$ there are already classification errors. This happens due to the occurrence of over-fitting for such high values of the parameter p.

4.2 Gaussian RBF kernel

Table 2

σ	Classification error (%)	Number of support vectors
2	0.07	81
1.5	0.01	74
0.9	0	56
0.89	0	52
0.8	0	54
0.7	0	66
0.5	0	82
0.1	0	100

Best value of σ = 0.89, Number of support vectors = 52

In table 2, we can find that as σ is diminished, there are found classifiers that are completely accurate. However, if σ keeps getting smaller, it's noticeable an increase in the number of support vectors. This occurs because a smaller value of σ tends to make a local classifier, therefore increasing the number of support vectors needed. In the opposite case, with larget values for σ, it tends to make a much more general classifier.

4.3 Best value of σ = 1, Number of support vectors = 10, plotted in Figure 2

4.4 Best value of σ = 1, Number of support vectors = 17, plotted in Figure 3

Since we are in the presence of an hard margin SVM, the classifier will compute a higher complexity model, as observable in figures 2 and 3.

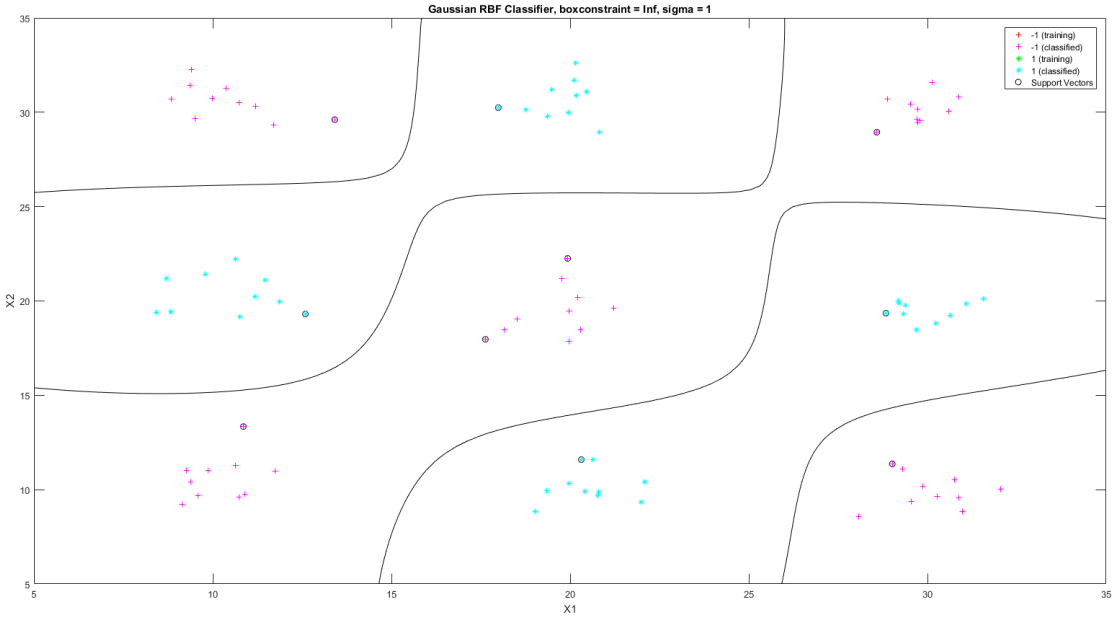


Figure 2

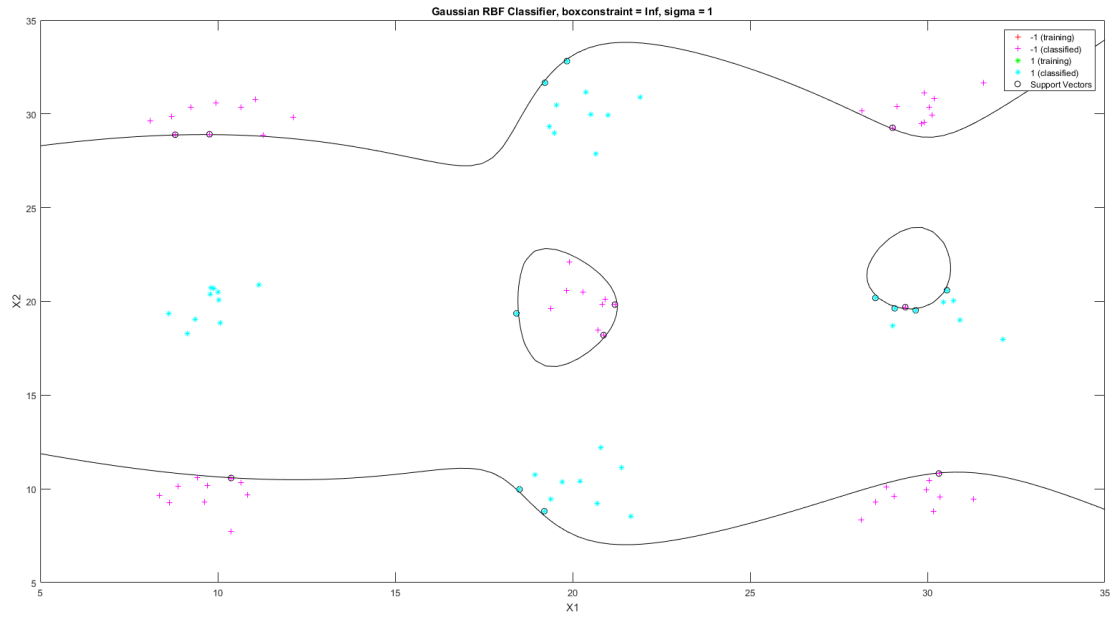


Figure 3

With outlier patterns, the number of support vectors rises, so that the classifier keeps it's accuracy, since we have an hard-margin SVM. In figure 3, two blobs are observable near the outlier points. The boundaries of these blobs have very small margins between the closest data points of classes '+1' and '-1'. This occurs due to overfitting of the classifier.

4.5

The soft margin consists on the idea of assigning a slack variable ξ_i to each data point x_i, and if $\xi_i > 0$ the point is on the wrong side of the hyperplane. With this said, it will be considered as a soft margin SVM the classifiers that haves errors in their classifications.

Table 3

'boxconstraint'	Classification error (%)	Number of support vectors
10.^7	0	16
10.^6	1	16
10.^5	1	18
10.^4	1	19
10.^3	1	20
10.^2	2	26
10.^1	2	62
10.^0	2	90

Diminishing the value of 'boxconstraint' gives a trade-off. For one hand, we diminish the over-fit of the data, by increasing the margin with a slack, decreasing the complexity of the model. On the other hand, training data will be more frequently classified incorrectly, and the number of support vectors needed increases. This is observable in table 3.