Instituto Superior Técnico

Departamento de Engenharia Electrotécnica e de Computadores

Machine Learning

5th Lab Assignment

Shift: Tuesday Group number: 2

Number: 81013 Name: José António Costa Coelho

Number: 81398 Name: Maria Carolina Varandas Roque

Support Vector Machines for Classification

1 Introduction

Simple linear classifiers, such as the one implemented by the Rosenblatt perceptron, are unable to correctly classify patterns, unless the classes under consideration are linearly separable. Neural networks that use hidden units with nonlinear activation functions are used in many classification problems, since they are able to perform nonlinear classification. However, several strong theoretical results, valid for the linearly separable case, are not applicable to nonlinear classifiers.

Support vector machines (SVMs) address the classification problem using linearly separable classes, not in the input space, but in the so-called *feature space*. Input patterns are mapped onto the higher-dimensional feature space, where the classification is performed using a hyperplane as classification border. Since the mapping from the input space to the feature space is usually nonlinear, these hyperplanes in feature space correspond to nonlinear borders in input space.

At first glance this might seem to be a double-edged sword, since it suggests that calculations have to be performed in the high-dimensional feature space. However, an interesting result proves that, since linear classification only requires inner product operations, all calculations can be performed in the lower-dimensional input space, if the nonlinear mapping is chosen in an appropriate way. This result is particularly strong when one takes into account that certain mappings yield infinite-dimensional feature spaces. This is the same as saying that linear classification in an infinite-dimensional feature space can be performed by means of operations in the lower-dimensional input space. Imagine all the power of infinite-dimensional hyperplanes, without the associated computational burden.

The purpose of this assignment is twofold: first, to work out, in detail, two simple classification problems in two-dimensional input space, one of them involving a mapping to a three-dimensional feature space; second, to provide some experience and some intuition on the capabilities of support vector machines.

2 Two simple examples

Consider the AND and XOR logic functions, defined in the following truth table:

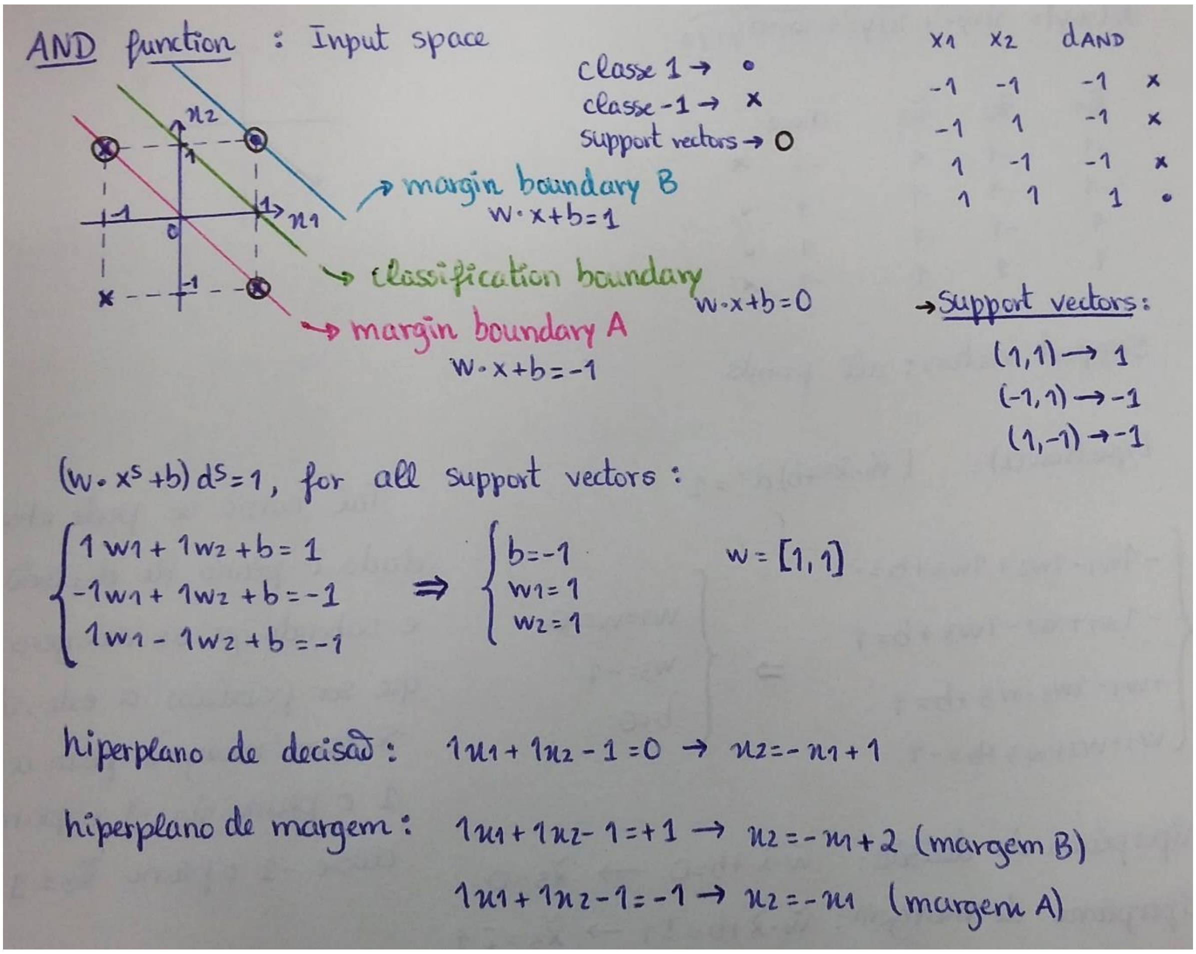
x_1	x_2	d_{AND}	d_{XOR}
-1	-1	-1	-1
-1	1	-1	1
1	-1	-1	1
1	1	1	-1

Here, the input pattern is a vector $\mathbf{x} = (x_1, x_2)$, and d_{AND} and d_{XOR} are the desired values for the AND and XOR functions. Note that, in this assignment, we represent logical *true* by 1 and logical *false* by −1. Similarly, in binary classification problems, we assign the desired value of 1 to the patterns of one of the classes, and the desired value of −1 to those of the other class.

2.1(T) For the AND function, find (by inspection) the maximum-margin separating straight line, the support vectors and the margin boundaries. Then compute the vector **w** and the bias *b* that satisfy the equation

$$(\mathbf{w} \cdot \mathbf{x}^s + b)d^s = 1 \tag{1}$$

for all support vectors $\mathbf{x}^s$, where d^s is the desired value corresponding to $\mathbf{x}^s$.[1]



AND function : Input space

classe 1 → •
classe -1 → ×
Support vectors → O

→ margin boundary B
$w \cdot x + b = 1$

→ classification boundary
$w \cdot x + b = 0$

→ margin boundary A
$w \cdot x + b = -1$

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

→ Support vectors:

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

$(w \cdot x^s + b)\, d^s = 1$, for all support vectors :

$$\begin{cases} 1w_1 + 1w_2 + b = 1 \\ -1w_1 + 1w_2 + b = -1 \\ 1w_1 - 1w_2 + b = -1 \end{cases} \Rightarrow \begin{cases} b = -1 \\ w_1 = 1 \\ w_2 = 1 \end{cases} \qquad w = [1, 1]$$

hiperplano de decisão : $1x_1 + 1x_2 - 1 = 0 \rightarrow x_2 = -x_1 + 1$

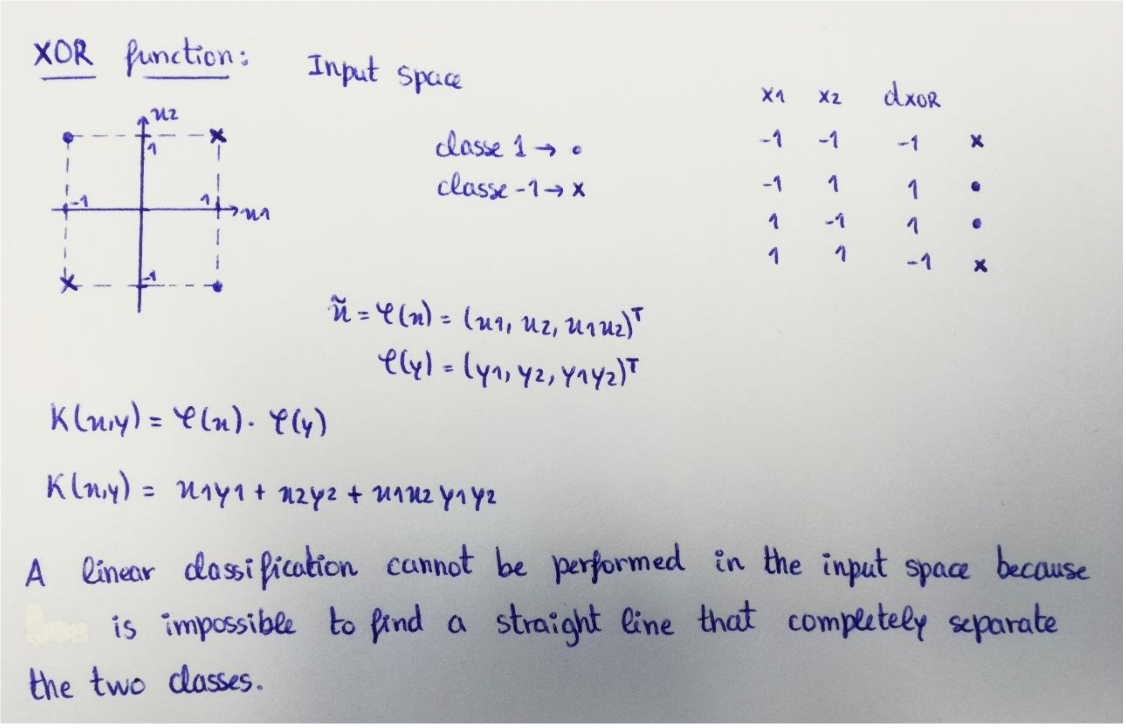
hiperplano de margem : $1x_1 + 1x_2 - 1 = +1 \rightarrow x_2 = -x_1 + 2$ (margem B)

$1x_1 + 1x_2 - 1 = -1 \rightarrow x_2 = -x_1$ (margem A)

2.2(T) Since, for the XOR function, a linear classification cannot be performed in the input space – explain why – we will consider here a simple nonlinear mapping to a three-dimensional feature space:

$$\tilde{\mathbf{x}} = \phi(\mathbf{x}) = (x_1, x_2, x_1 x_2)^T. \tag{4}$$

Find the kernel function that corresponds to this mapping.



XOR function: Input space

classe 1 → •
classe -1 → x

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

$$\tilde{x} = \varphi(x) = (x_1, x_2, x_1 x_2)^T$$

$$\varphi(y) = (y_1, y_2, y_1 y_2)^T$$

$$K(x,y) = \varphi(x) \cdot \varphi(y)$$

$$K(x,y) = x_1 y_1 + x_2 y_2 + x_1 x_2 y_1 y_2$$

A linear classification cannot be performed in the input space because is impossible to find a straight line that completely separate the two classes.

[1] It can be easily shown (but you're not asked to show) that, defining the border of the maximum-margin linear classifier by the equation

$$\mathbf{w} \cdot \tilde{\mathbf{x}} + b = 0, \tag{2}$$

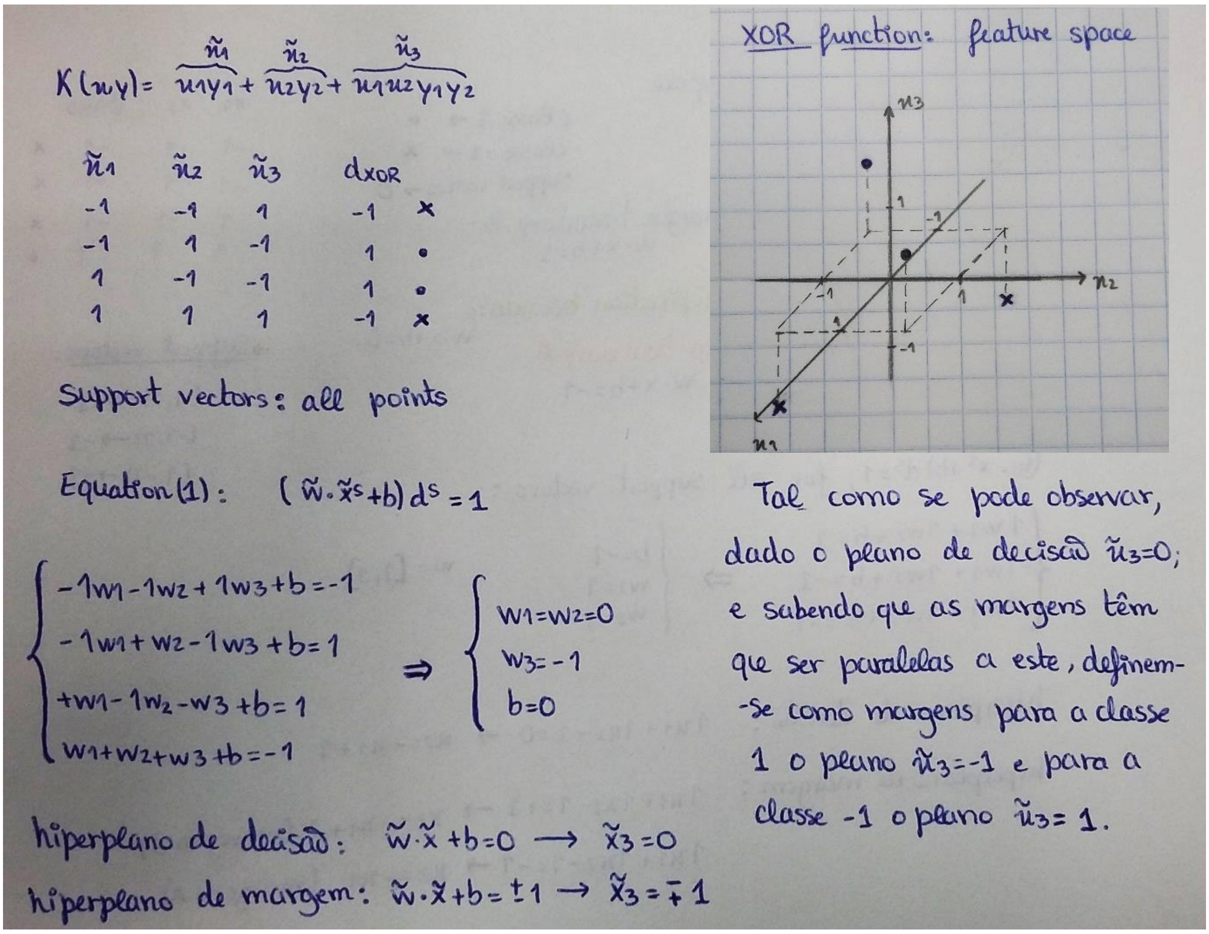
then $\mathbf{w}$ and b obey the equation

$$(\mathbf{w} \cdot \tilde{\mathbf{x}}^s + b) d^s = C \tag{3}$$

for all support vectors $\tilde{\mathbf{x}}^s$, where C is a constant. Normally, we choose $C = 1$.

In our case of the AND function, vectors with and without tilde are equal, since the feature space (where the linear classification is performed) is the input space itself.

2.3(T) Visualize the points in this 3D feature space. Find, by inspection, which are the support vectors. Compute **w** and *b* in this feature space, so e that equation (1) is satisfied for all support vectors.



$K(x,y) = \overbrace{x_1y_1}^{\tilde{x}_1} + \overbrace{x_2y_2}^{\tilde{x}_2} + \overbrace{x_1x_2y_1y_2}^{\tilde{x}_3}$

$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	d_{XOR}	
-1	-1	1	-1	×
-1	1	-1	1	●
1	-1	-1	1	●
1	1	1	-1	×

Support vectors: all points

Equation (1): $(\tilde{w} \cdot \tilde{x}^s + b)\, d^s = 1$

$$\begin{cases} -1w_1 - 1w_2 + 1w_3 + b = -1 \\ -1w_1 + w_2 - 1w_3 + b = 1 \\ +w_1 - 1w_2 - w_3 + b = 1 \\ w_1 + w_2 + w_3 + b = -1 \end{cases} \Rightarrow \begin{cases} w_1 = w_2 = 0 \\ w_3 = -1 \\ b = 0 \end{cases}$$

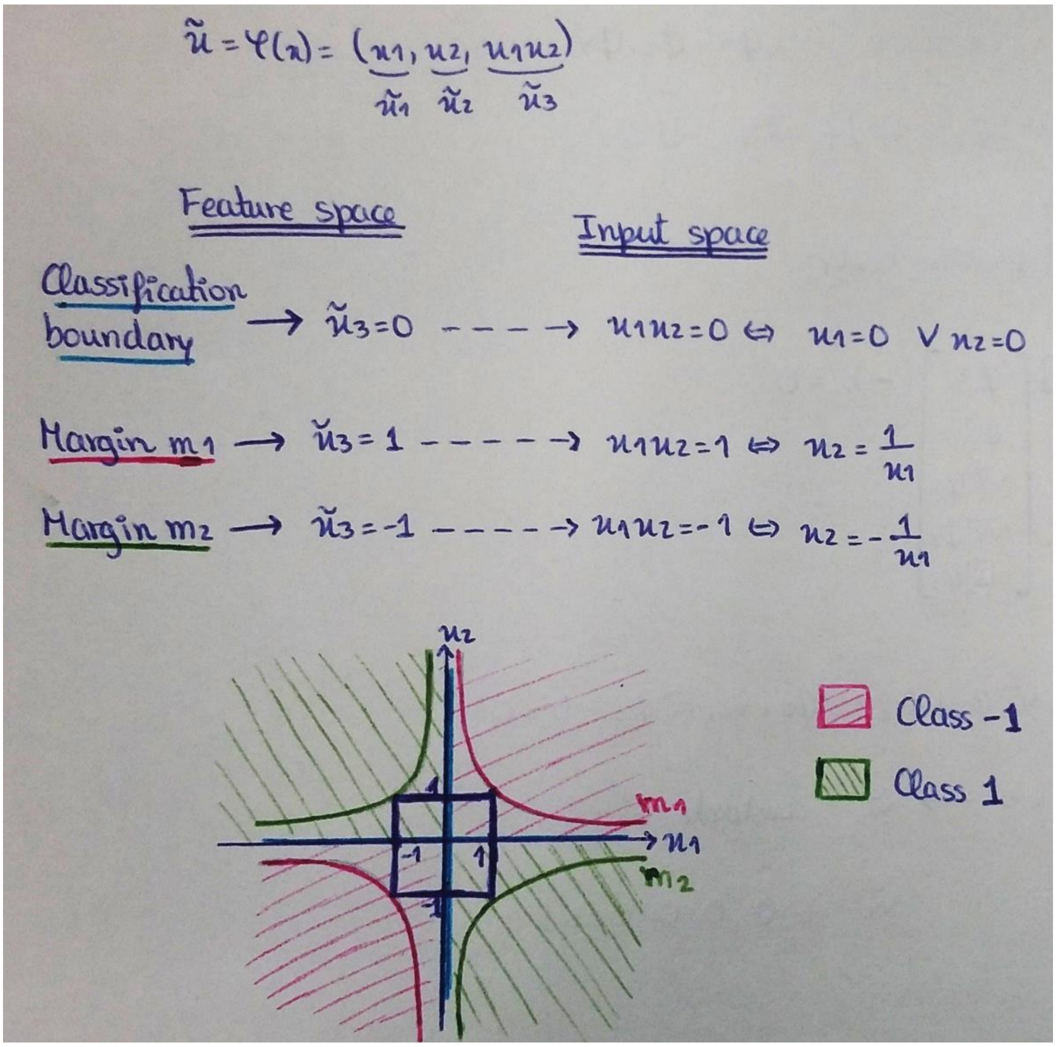
hiperplano de decisão: $\tilde{w} \cdot \tilde{x} + b = 0 \longrightarrow \tilde{x}_3 = 0$

hiperplano de margem: $\tilde{w} \cdot \tilde{x} + b = \pm 1 \rightarrow \tilde{x}_3 = \mp 1$

XOR function: feature space

Tal como se pode observar, dado o plano de decisão $\tilde{x}_3 = 0$; e sabendo que as margens têm que ser paralelas a este, definem-se como margens para a classe 1 o plano $\tilde{x}_3 = -1$ e para a classe -1 o plano $\tilde{x}_3 = 1$.

2.4(T) Algebraically express, in the two-dimensional input space, the classification border and the margin boundaries corresponding to the classifier found above for the XOR problem. Then sketch them in a graph, together with the input patterns.



$$\tilde{x} = \varphi(x) = (\underbrace{x_1}_{\tilde{x}_1}, \underbrace{x_2}_{\tilde{x}_2}, \underbrace{x_1x_2}_{\tilde{x}_3})$$

Feature space

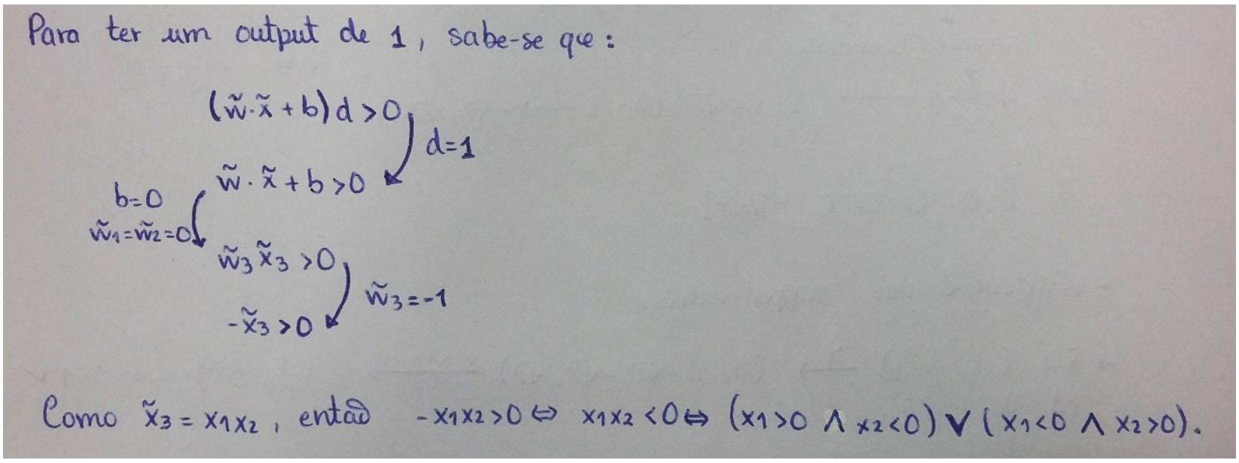
Input space

Classification boundary → $\tilde{x}_3=0$ ----→ $x_1x_2=0 \Leftrightarrow x_1=0 \vee x_2=0$

Margin m1 → $\tilde{x}_3=1$ ----→ $x_1x_2=1 \Leftrightarrow x_2=\frac{1}{x_1}$

Margin m2 → $\tilde{x}_3=-1$ ----→ $x_1x_2=-1 \Leftrightarrow x_2=-\frac{1}{x_1}$

2.5(T) Indicate the mathematical condition under which the classifier that you have just developed will produce an output of 1. The condition should be expressed in terms of the input space coordinates. It shouldn't use coordinates from the feature space.



Para ter um output de 1, sabe-se que:

$$(\tilde{w}\cdot\tilde{x}+b)d>0 \xrightarrow{d=1} \tilde{w}\cdot\tilde{x}+b>0 \xrightarrow{b=0,\ \tilde{w}_1=\tilde{w}_2=0} \tilde{w}_3\tilde{x}_3>0 \xrightarrow{\tilde{w}_3=-1} -\tilde{x}_3>0$$

Como $\tilde{x}_3 = x_1x_2$, então $-x_1x_2>0 \Leftrightarrow x_1x_2<0 \Leftrightarrow (x_1>0 \wedge x_2<0) \vee (x_1<0 \wedge x_2>0)$.

3 Classification using SVMs

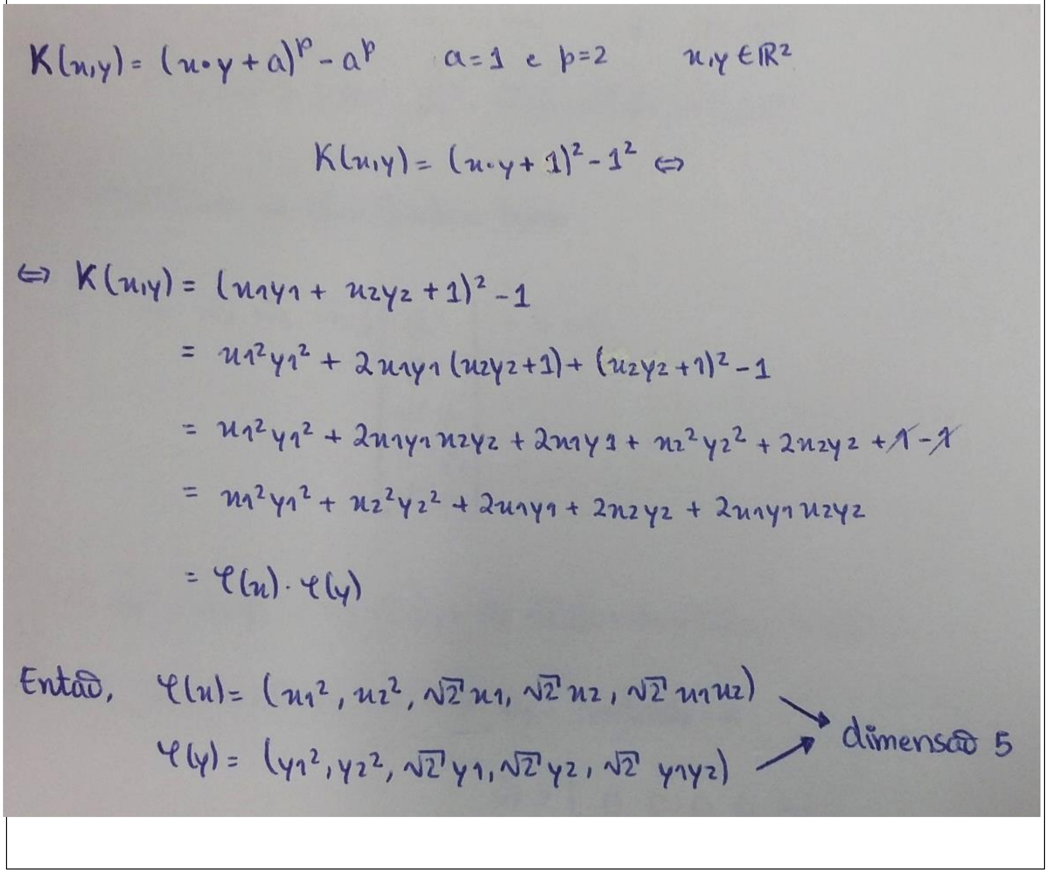
A kernel commonly employed in pattern recognition problems is the polynomial one, defined by

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + a)^p - a^p, \qquad (5)$$

where $a \in$ R⁺ and $p \in$ N.[1]

3.1(T) Consider $\mathbf{x},\mathbf{y} \in$ R², a = 1 and p = 2. Indicate the mapping to feature space that this kernel corresponds to, and the dimensionality of the feature space.

R.



$K(x,y) = (x \cdot y + a)^p - a^p \qquad a = 1$ e $p = 2 \qquad x,y \in \mathbb{R}^2$

$$K(x,y) = (x \cdot y + 1)^2 - 1^2 \Leftrightarrow$$

$$\Leftrightarrow K(x,y) = (x_1y_1 + x_2y_2 + 1)^2 - 1$$

$$= x_1^2y_1^2 + 2x_1y_1(x_2y_2+1) + (x_2y_2+1)^2 - 1$$

$$= x_1^2y_1^2 + 2x_1y_1x_2y_2 + 2x_1y_1 + x_2^2y_2^2 + 2x_2y_2 + \cancel{1} - \cancel{1}$$

$$= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 + 2x_1y_1x_2y_2$$

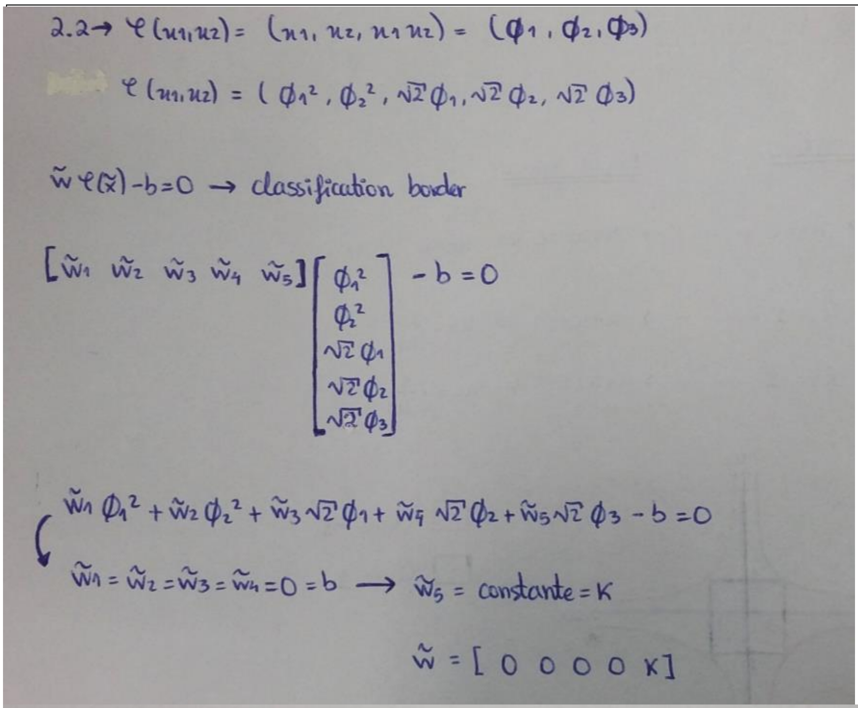
$$= \varphi(x) \cdot \varphi(y)$$

Então, $\varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2) \searrow$

$\varphi(y) = (y_1^2, y_2^2, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}\,y_1y_2) \nearrow$ dimensão 5

[1] This is one of the variants of the polynomial kernel. Another variant omits the term "$-a^p$" in the defining equation.

3.2(T) Assume again that p = 2 and a = 1. Find the vector **w** that repree sents, in this new feature space, the same classification border and margins as in 2.2.



$2.2 \rightarrow \varphi(x_1, x_2) = (x_1, x_2, x_1 x_2) = (\phi_1, \phi_2, \phi_3)$

$\varphi(x_1, x_2) = (\phi_1^2, \phi_2^2, \sqrt{2}\phi_1, \sqrt{2}\phi_2, \sqrt{2}\phi_3)$

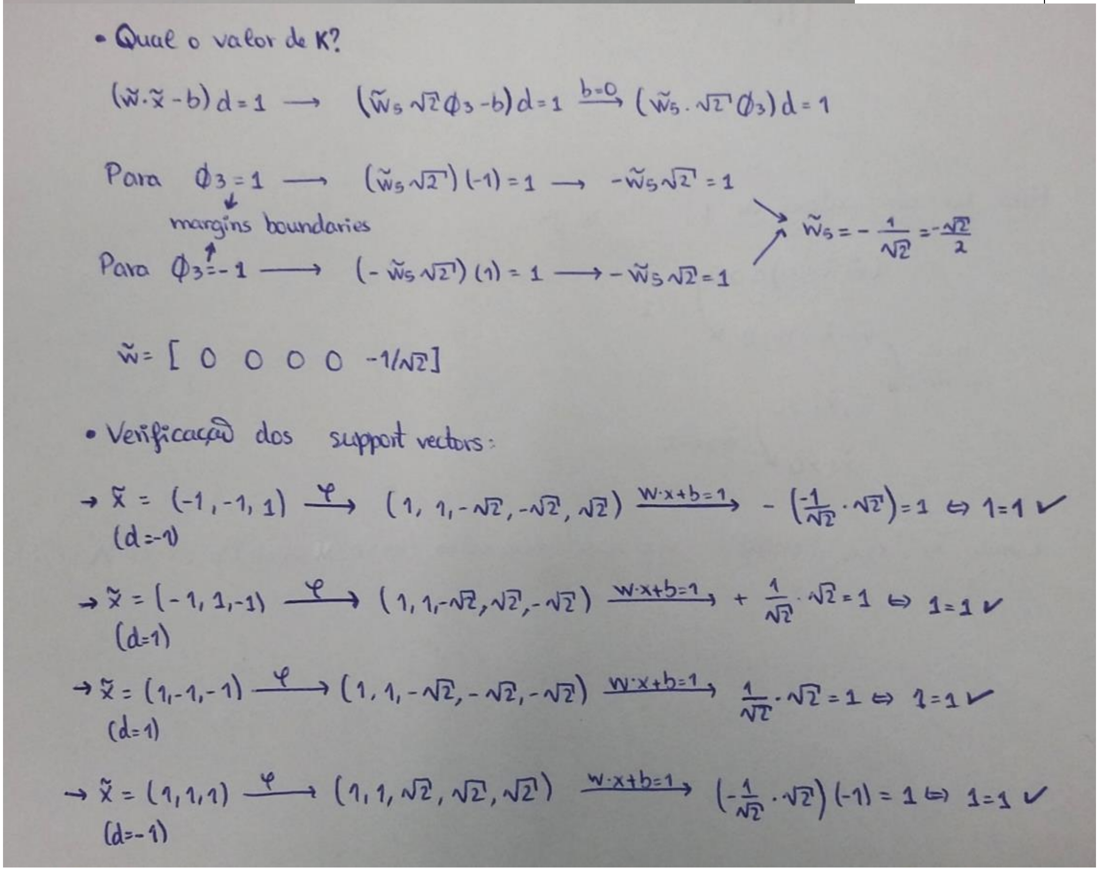
$\tilde{w}\varphi(\tilde{x}) - b = 0 \rightarrow$ classification border

$$[\tilde{w}_1\ \tilde{w}_2\ \tilde{w}_3\ \tilde{w}_4\ \tilde{w}_5]\begin{bmatrix}\phi_1^2\\ \phi_2^2\\ \sqrt{2}\phi_1\\ \sqrt{2}\phi_2\\ \sqrt{2}\phi_3\end{bmatrix} - b = 0$$

$\tilde{w}_1\phi_1^2 + \tilde{w}_2\phi_2^2 + \tilde{w}_3\sqrt{2}\phi_1 + \tilde{w}_4\sqrt{2}\phi_2 + \tilde{w}_5\sqrt{2}\phi_3 - b = 0$

$\tilde{w}_1 = \tilde{w}_2 = \tilde{w}_3 = \tilde{w}_4 = 0 = b \longrightarrow \tilde{w}_5 =$ constante $= K$

$\tilde{w} = [\,0\ \ 0\ \ 0\ \ 0\ \ K\,]$



- Qual o valor de K?

$(\tilde{w}\cdot\tilde{x} - b)d = 1 \longrightarrow (\tilde{w}_5\sqrt{2}\phi_3 - b)d = 1 \xrightarrow{b=0} (\tilde{w}_5 \cdot \sqrt{2}\phi_3)d = 1$

Para $\phi_3 = 1 \longrightarrow (\tilde{w}_5\sqrt{2})(-1) = 1 \longrightarrow -\tilde{w}_5\sqrt{2} = 1$

margins boundaries

Para $\phi_3 = -1 \longrightarrow (-\tilde{w}_5\sqrt{2})(1) = 1 \longrightarrow -\tilde{w}_5\sqrt{2} = 1$

$\tilde{w}_5 = -\frac{1}{\sqrt{2}} = -\frac{\sqrt{2}}{2}$

$\tilde{w} = [\,0\ \ 0\ \ 0\ \ 0\ \ -1/\sqrt{2}\,]$

- Verificação dos support vectors:

$\rightarrow \tilde{x} = (-1, -1, 1) \xrightarrow{\varphi} (1, 1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}) \xrightarrow{w\cdot x + b = 1} -\left(\frac{-1}{\sqrt{2}} \cdot \sqrt{2}\right) = 1 \Leftrightarrow 1 = 1$ ✓
$(d = -1)$

$\rightarrow \tilde{x} = (-1, 1, -1) \xrightarrow{\varphi} (1, 1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}) \xrightarrow{w\cdot x + b = 1} + \frac{1}{\sqrt{2}} \cdot \sqrt{2} = 1 \Leftrightarrow 1 = 1$ ✓
$(d = 1)$

$\rightarrow \tilde{x} = (1, -1, -1) \xrightarrow{\varphi} (1, 1, -\sqrt{2}, -\sqrt{2}, -\sqrt{2}) \xrightarrow{w\cdot x + b = 1} \frac{1}{\sqrt{2}} \cdot \sqrt{2} = 1 \Leftrightarrow 1 = 1$ ✓
$(d = 1)$

$\rightarrow \tilde{x} = (1, 1, 1) \xrightarrow{\varphi} (1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2}) \xrightarrow{w\cdot x + b = 1} \left(-\frac{1}{\sqrt{2}} \cdot \sqrt{2}\right)(-1) = 1 \Leftrightarrow 1 = 1$ ✓
$(d = -1)$

4 Experiments

The experimental part of this assignment uses the SVM toolbox from MatLab. Use function svmtrain for training the SVM and function svmclassify for testing (type help for more information on these functions). You will need to specify the ’kernel _function’. Use ’linear’ for a linear classifier (*i.e.,* the feature space is equal to the input space), ’polynomial’ for the kernel (5), where ’polyorder’ stands for parameter *p*, and ’rbf’ (radial basis function) for the kernel

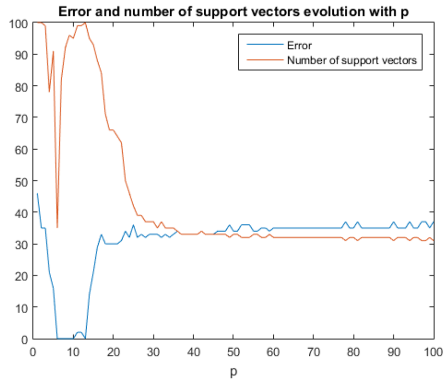
$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}, \tag{6}$$

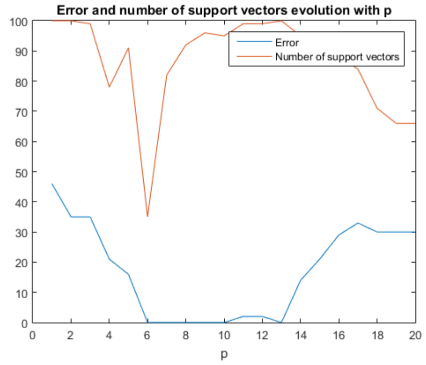
where ’rbf _sigma’ stands for σ. Among these kernels, Gaussian RBF is the one that is most frequently used, for several reasons (for instance, because it is shift-invariant and isotropic).

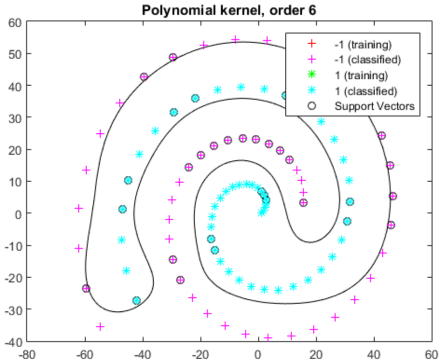
Set the svmtrain parameter ’Method’ to ’QP’ to choose Quadratic Programming as the optimization method, and the ’boxconstraint’ parameter to 10^4. This parameter corresponds to the soft margin penalty ’C’ which specifies the relative weight of the margin violations in the objective function that is optimized in the training of the classifier.

When training and testing your classifiers, set option ’Showplot’ to true in order to obtain the plot of the classification.

4.1(E) Load the file spiral.mat. This file contains the classical spiral example, with 50 patterns per class. Determine experimentally, using the polynomial kernel, the value of *p* for which you get the best classifier. (start with *p* = 1). Write down all experiments performed, together with the classification error percentages and number of support vectors (the support vectors can be obtained from the SVMStruct returned by svmtrain). Comment on the results you obtained.



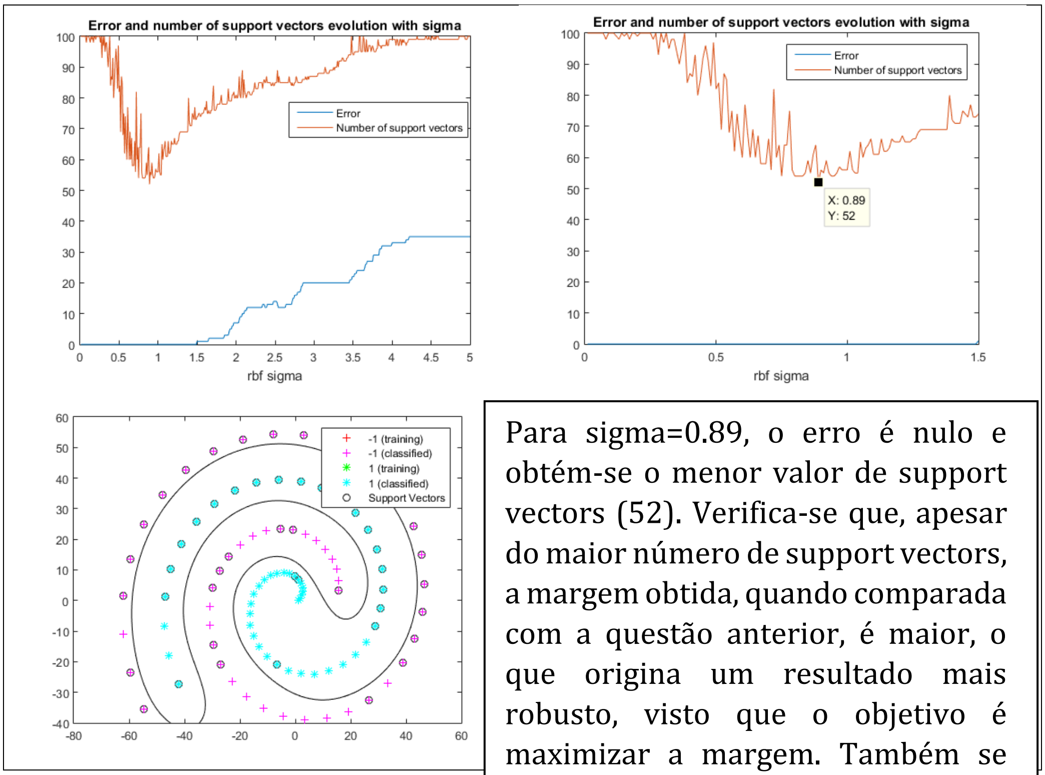




Para p=6, o erro é mínimo e obtém-se o menor valor de support vectors. Contudo, é possível ver que a margem é muito baixa (por comparação da Figura ao lado, na qual está indicado o plano de separação das classes e os support vectors.

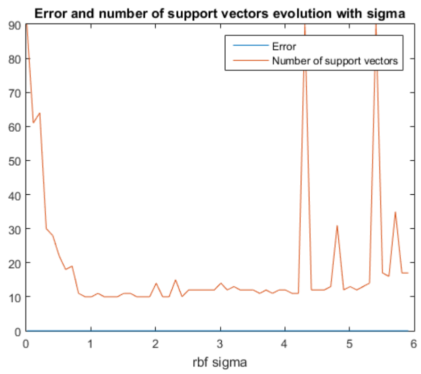
Verifica-se que não há vantagem em utilizar um valor de p muito elevado, dado que isso origina um erro de aproximadamente 30%, como se vê no gráfico acima.

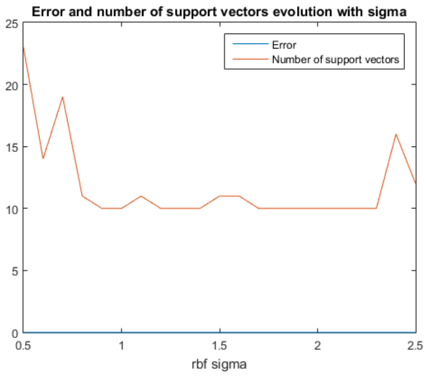
4.2(E) Using the same data file (spiral.mat), try now the Gaussian RBF kernel. Find the approximate value of σ for which you can get the best classifier. Comment on the results you obtained.



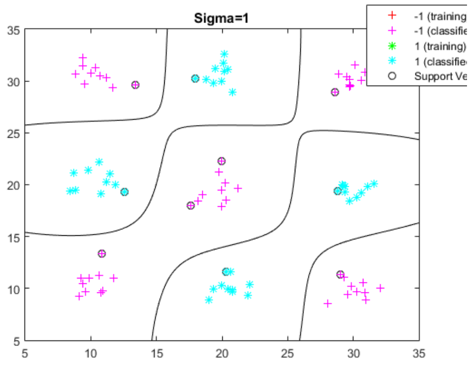
Para sigma=0.89, o erro é nulo e obtém-se o menor valor de support vectors (52). Verifica-se que, apesar do maior número de support vectors, a margem obtida, quando comparada com a questão anterior, é maior, o que origina um resultado mais robusto, visto que o objetivo é maximizar a margem. Também se pode concluir que a utilização de um valor de sigma muito elevado origina a um resultado com mais erros e com a necessidade de recorrer a mais support vectors.

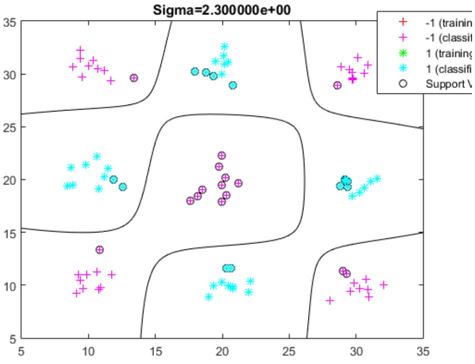
4.3(E) Load the file chess33.mat and set 'boxconstraint' parameter to Inf to enforce a hard margin SVM, for separable data. Using the Gaussian RBF kernel, find a value of σ that approximately minimizes the number of support vectors, while correctly classifying all patterns. Indicate the value of σ and the number of support vectors.





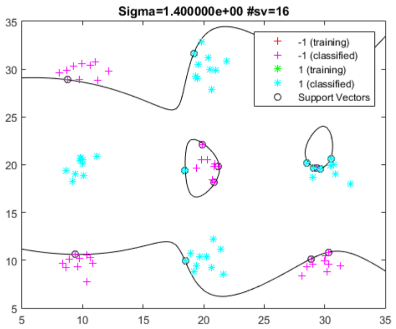
Pode-se ver que há um conjunto de valores de sigma que originam uma utilização mínima de support vectors. Dado o objetivo de maximizar a margem, observaram-se os plots para os vários valores de sigma, e escolheu-se o valor que origina margens maiores.

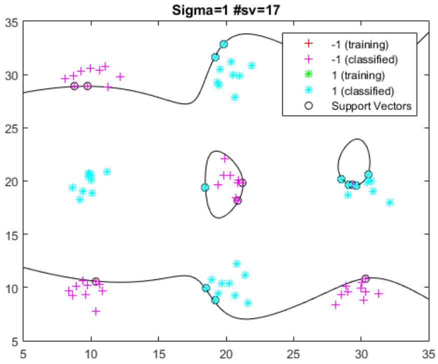




Como se pode observar, o valor de sigma=2.3 origina maiores margens (maior distancia dos support vectors à fronteira), pelo que se pode considerar este como o valor optimo.

4.4(E) Load the file chess33n.mat which is similar to the one used in the previous question, except for the presence of a couple of outlier patterns. Run the classification algorithm on these data with the same value of σ, and comment on how the results changed, including the shape of the classification border, the margin size and the number of support vectors.





Neste caso, não foi possível utilizar o valor de sigma obtido anteriormente. Em alternativa, procurou-se o valor que maximiza a margem, tendo sido, para isso, escolhido o valor de sigma=1.4