Aprendizagem de Máquina 1

Inteligência Artificial

PECE Programa de Educação Continuada
Escola Politécnica da USP

USP

AULA 05 – MÁQUINAS DE VETORES DE SUPORTE

Larissa Driemeier
Thiago Martins

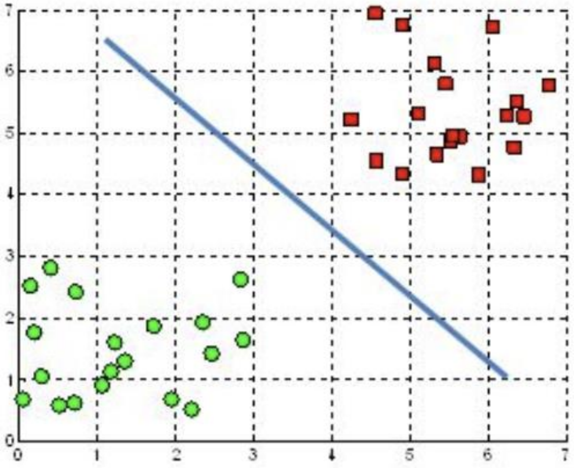
PECE Programa de Educação Continuada
Escola Politécnica da USP

CRONOGRAMA

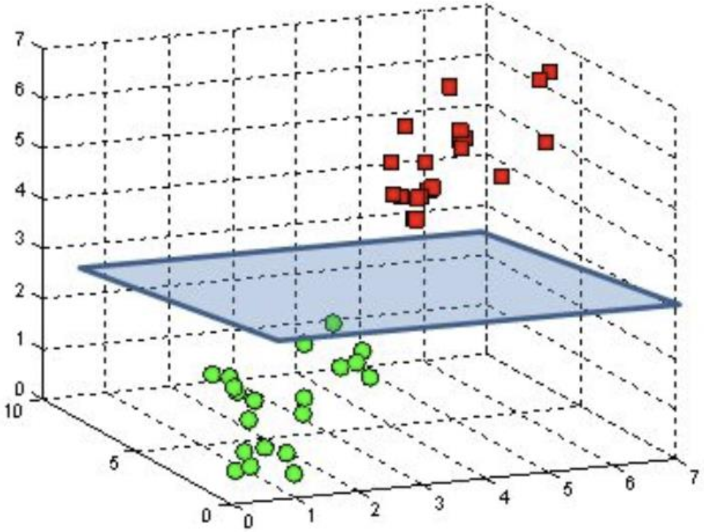
Data	Professor	Assunto
07/05	Larissa	Definição de aprendizado de máquina. Aprendizado supervisionado e não supervisionado. Regressão linear. Regressão polinomial.
14/05	Thiago	Exercícios de acompanhamento. Nota 01.
21/05	Larissa	Regressão Logística
28/05	Thiago	Exercícios de acompanhamento. Nota 02.
04/06	Larissa	Máquinas de vetores de suporte
11/06	Thiago	Exercícios de acompanhamento. Nota 03.
18/06	Larissa	Aprendizado não supervisionado
25/06	Thiago	Exercícios de acompanhamento. Nota 04.
02/07	Larissa	Redução de similaridade: análise de componentes principais (PCA) e suas variações.
16/07	Larissa/Thiago	Exercícios “Melhores Momentos”. Nota 05.

HIPERPLANO

- Em geometria, hiperplano é um subespaço de dimensão menor que o espaço de ambiente;
- Por exemplo, um hiperplano de um espaço de dimensão n é um subconjunto plano com dimensão $n - 1$.



Hiperplano no espaço $\mathcal{R}^2$ é uma linha



Hiperplano no espaço $\mathcal{R}^3$ é um plano

Simplificando, é como se o hiperplano **separasse o espaço em dois meio espaços...**

```python
SVM = SVC(kernel='rbf', gamma=.10, C=1.0)
SVM.fit(X_train_standard, y_train)
```

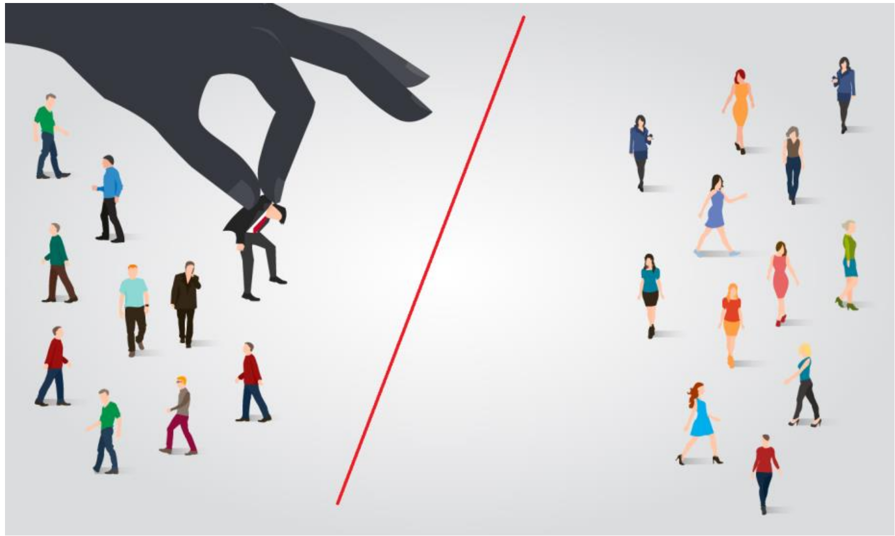
SVM (SUPPORT VECTOR MACHINE)

Máquinas de Vetores de Suporte

MÁQUINA DE VETORES DE SUPORTE

SVM (Support Vector Machine) é um classificador formalmente definido por um hiperplano de separação. Em outras palavras, com dados de treinamento rotulados (aprendizado supervisionado), o algoritmo gera um hiperplano ideal que categoriza novos exemplos.

- Sem dúvida, é o método de classificação de maior sucesso em aprendizado de máquinas;
- O método tem uma derivação matemática consistente;
- Existem ótimas *ferramentas numéricas* que podem ser usados para resolver o problema;
- A solução tem uma interpretação geométrica bastante intuitiva.



COMO CHEGAR LÁ?

- Margens funcionais e geométricas
- Classificador de margem ótima
- Dualidade de Lagrange
- Regularização e não separabilidade
- Kernels e o truque de Kernel

Solução

Transformações não lineares

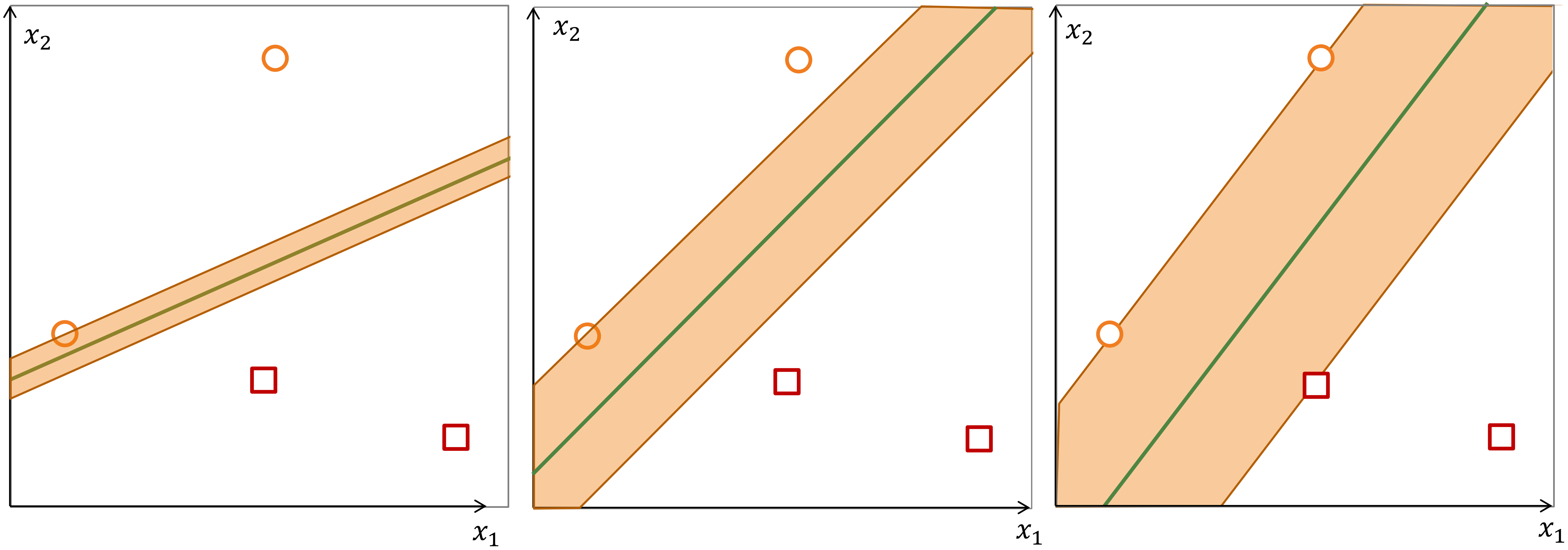
PECE Programa de Educação Continuada
Escola Politécnica da USP

MARGENS

Intuição

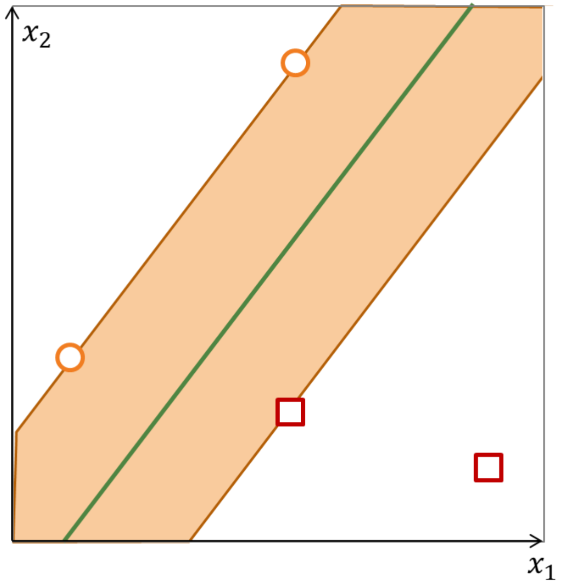
INTUIÇÃO...

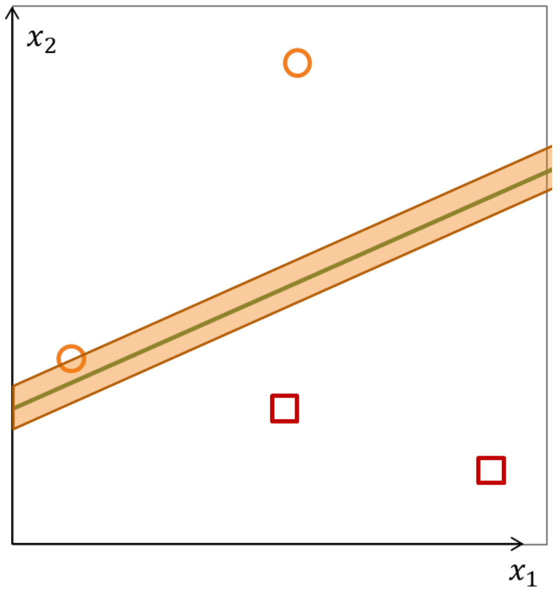
Qual a melhor separação para meu problema de classificação???



SURGEM DAÍ DUAS PERGUNTAS:

1. Porque a margem maior é melhor?
2. Se estamos convencidos de que a margem maior é melhor, então quais os pesos $\boldsymbol{\omega}$ que maximizam a margem?

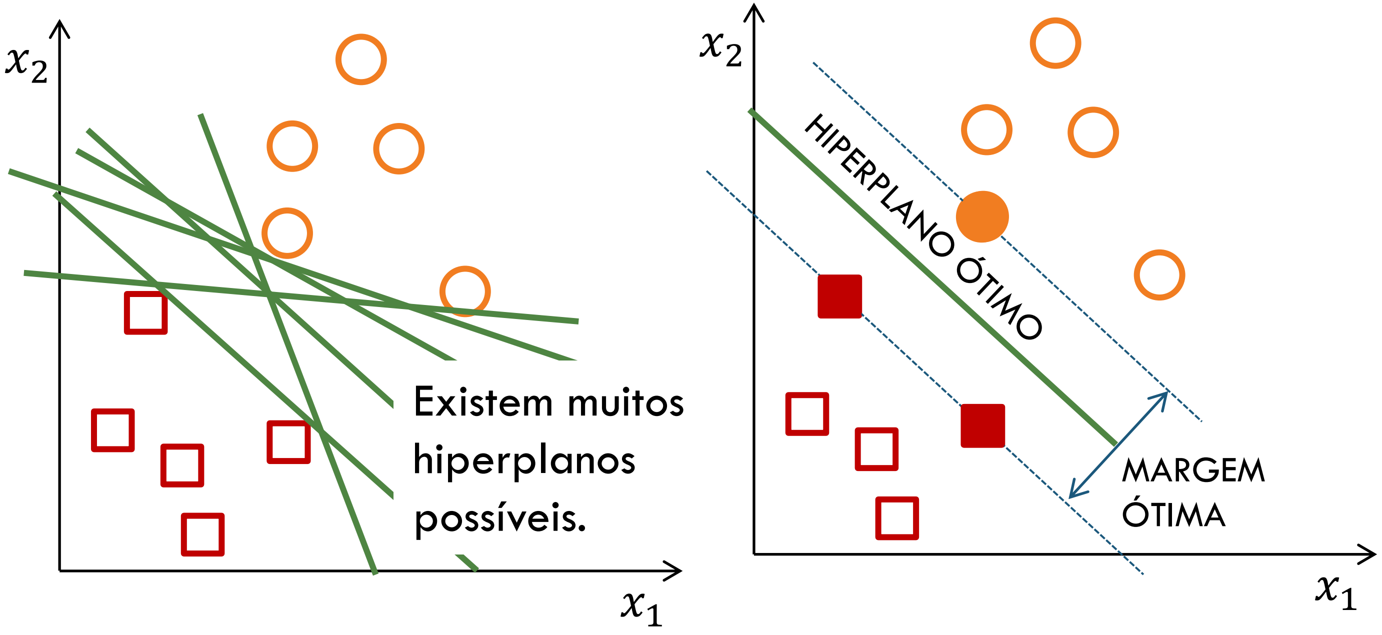




SUPPORT VECTOR MACHINE

Widest street approach

Nosso objetivo é encontrar um plano que tenha a margem máxima, ou seja, a distância máxima entre os pontos de dados das duas classes.



CONSIDERAÇÕES TÉCNICAS

Nos ajudarão **muito** futuramente

NOVA NOTAÇÃO

Dados de aprendizado $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$, onde $y^{(i)} \in (-1,1)$

$$\boldsymbol{\omega}^T\boldsymbol{x} + b = b + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n \qquad \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_n \end{bmatrix}$$

$$\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$$

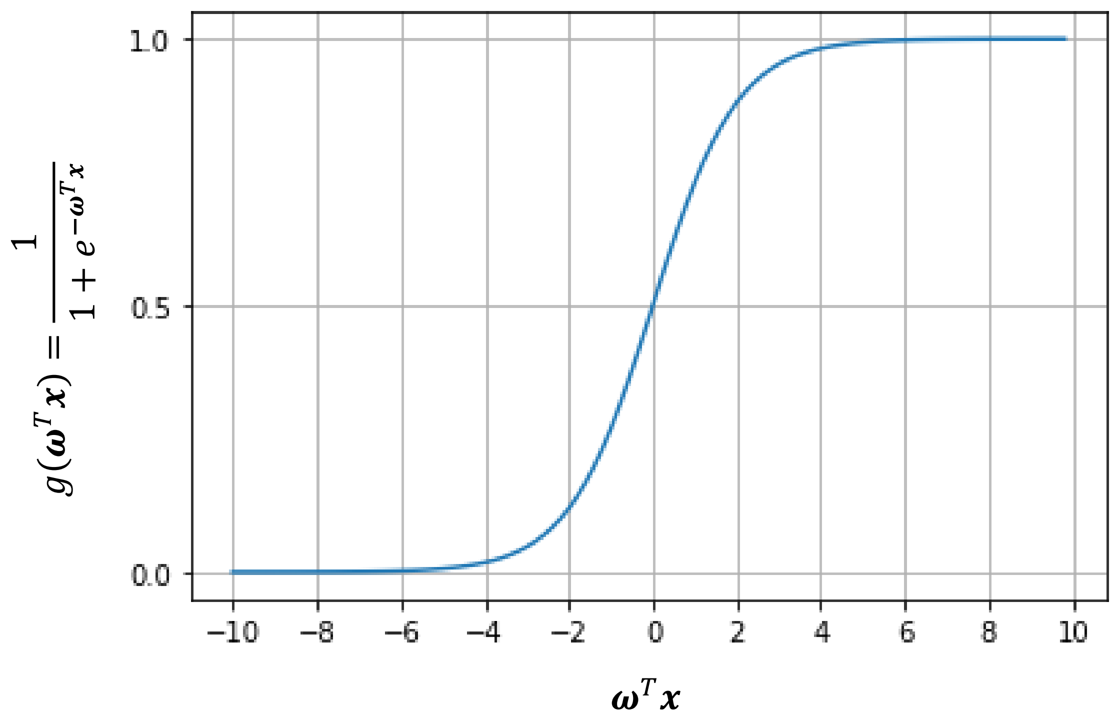
Veremos, mais adiante, que, com essa nova notação, nosso classificador irá predizer diretamente -1 ou $+1$, sem a etapa intermediária de regressão logística que estima a probabilidade de $y = 0$ ou $y = 1$.

LEMBREM-SE DE REGRESSÃO LOGÍSTICA

PECE Programa de Educação Continuada
Escola Politécnica da USP

$$h(\boldsymbol{x}) = g(\boldsymbol{\omega}^T \boldsymbol{x}) = \frac{1}{1+e^{-\omega^T x}}$$

Se $y^{(i)} = 1$ queremos $h(\boldsymbol{x}) \approx 1$, isto é $\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} \gg 0$
Se $y^{(i)} = 0$ queremos $h(\boldsymbol{x}) \approx 0$, isto é $\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} \ll 0$



Essa é uma boa ideia, e será a base para definição de ***margem***.

COMEÇANDO A ENTRAR NA TEORIA DA MARGEM...

A margem é somente a distância entre um ponto e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$

Se

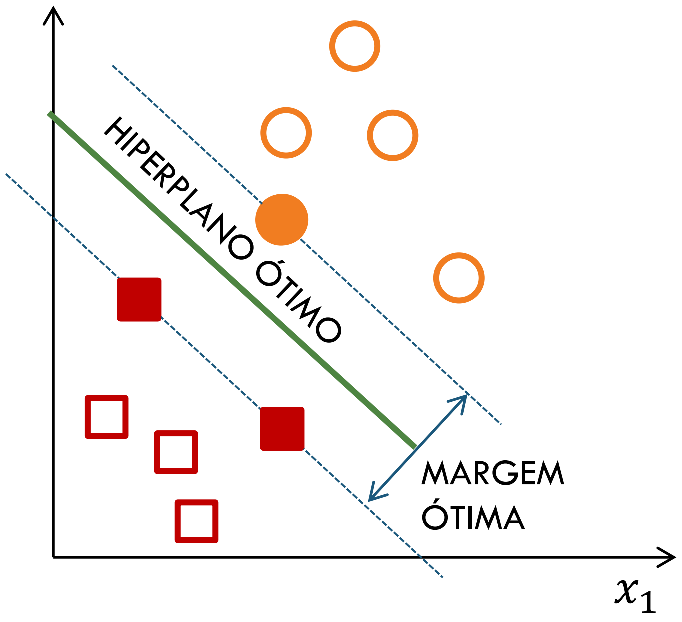
$y^{(i)} = 1$ então, para margem ser grande, $\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b \gg 0$

$y^{(i)} = -1$ então, para margem ser grande, $\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b \ll 0$

Então, se $y^{(i)} = \pm 1$

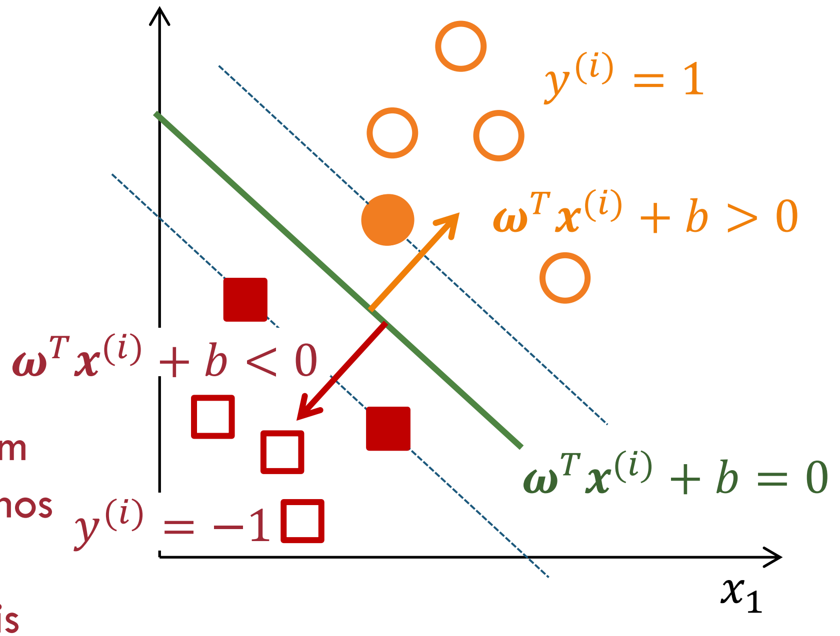
$$\hat{\gamma}^{(i)} = y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) = \left|\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right| > 0$$

Conceito de margem



$$\hat{\gamma}^{(i)} = y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = \left|\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right| > 0$$

PECE Programa de Educação Continuada
Escola Politécnica da USP



$\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b > 0$ para uma margem funcional quer dizer que estamos nos movendo em uma direção da reta, para cima

$\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b < 0$ para uma margem funcional quer dizer que estamos nos movendo para outra direção, encontramos o ponto negativo mais próximo e então os pontos interiores

$$\hat{\gamma}^{(i)} = y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = \left|\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right| > 0$$

Ou seja, se $\hat{\gamma}^{(i)} = y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) > 0$, a classificação está correta!

Na prática, essa hipótese $\hat{\gamma}^{(i)} > 0$ afirma que os nossos dados são (perfeitamente) linearmente separáveis.

SOBRE NOSSA NOVA HIPÓTESE

$$h(\boldsymbol{x}) = g(\boldsymbol{\omega}^T \boldsymbol{x} + b) = \mathrm{sign}(\boldsymbol{\omega}^T \boldsymbol{x} + b)$$

MAS...

Plano de divisão...

$$\cancel{1000}\ \boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + \cancel{1000}\, b = 0$$

Margem...

$$\hat{\gamma} = 1000\ y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right)$$

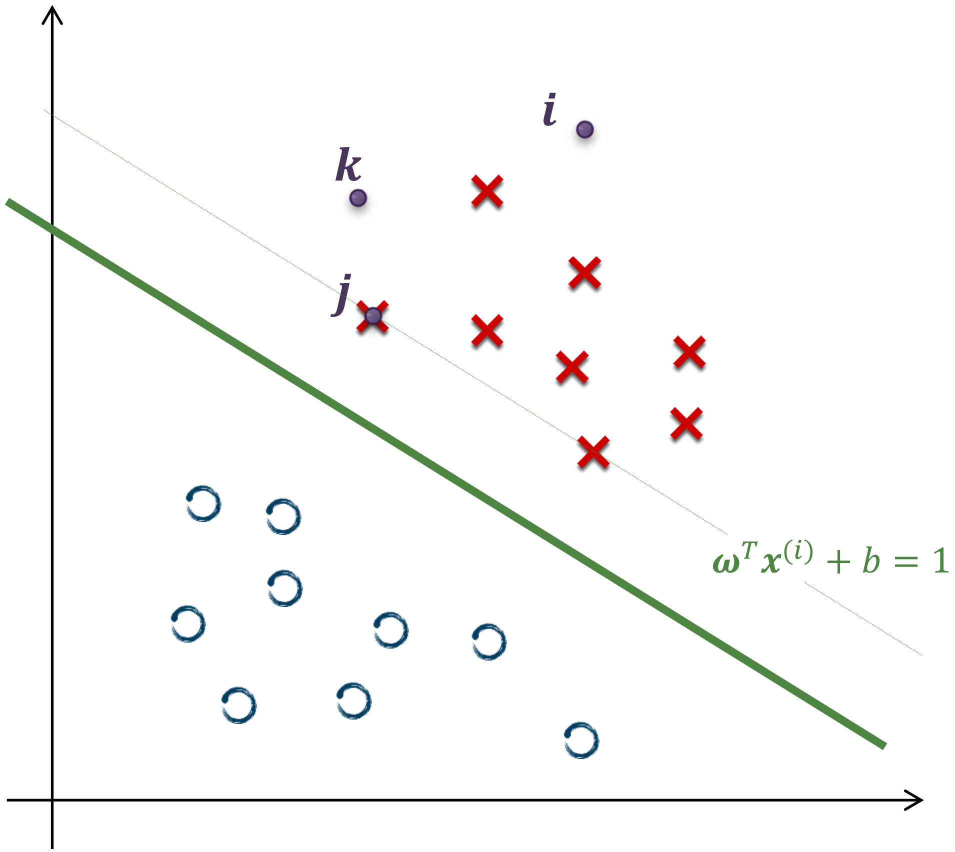
Achei uma maneira de deixar a margem tão grande quanto eu queira....

ENTÃO...

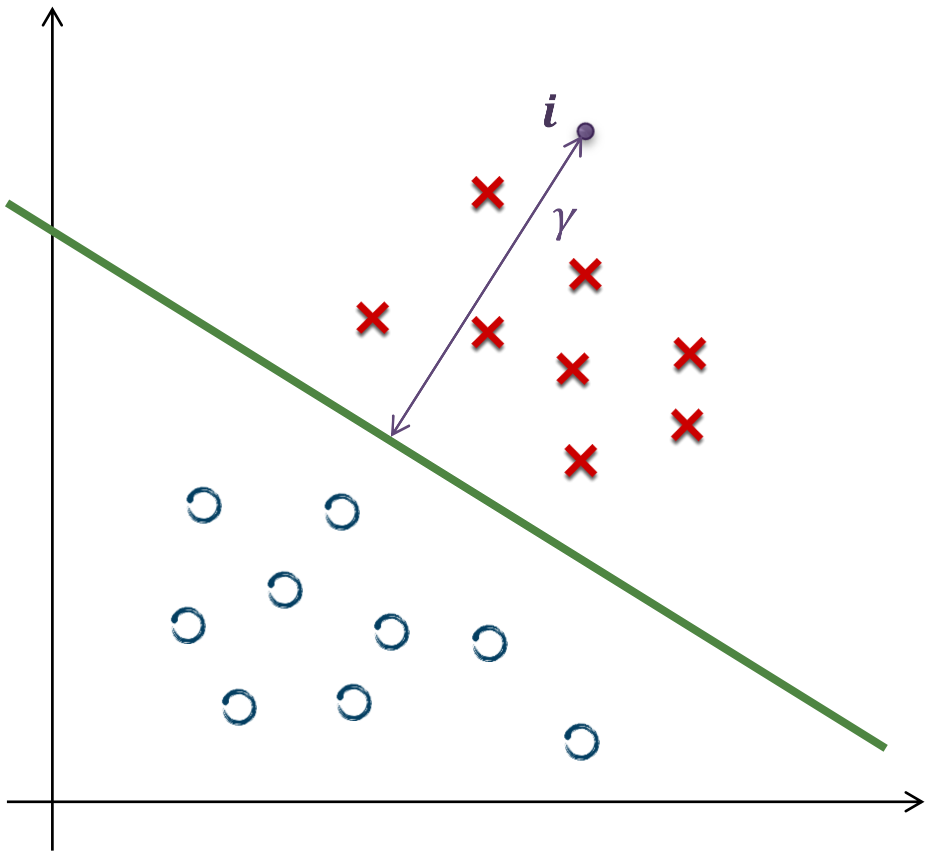
$$\hat{\gamma}^{(i)} = y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) > 0$$

$$\boxed{\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}}$$

$$\hat{\gamma} = \min_{\boldsymbol{\omega},b} y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) = 1$$



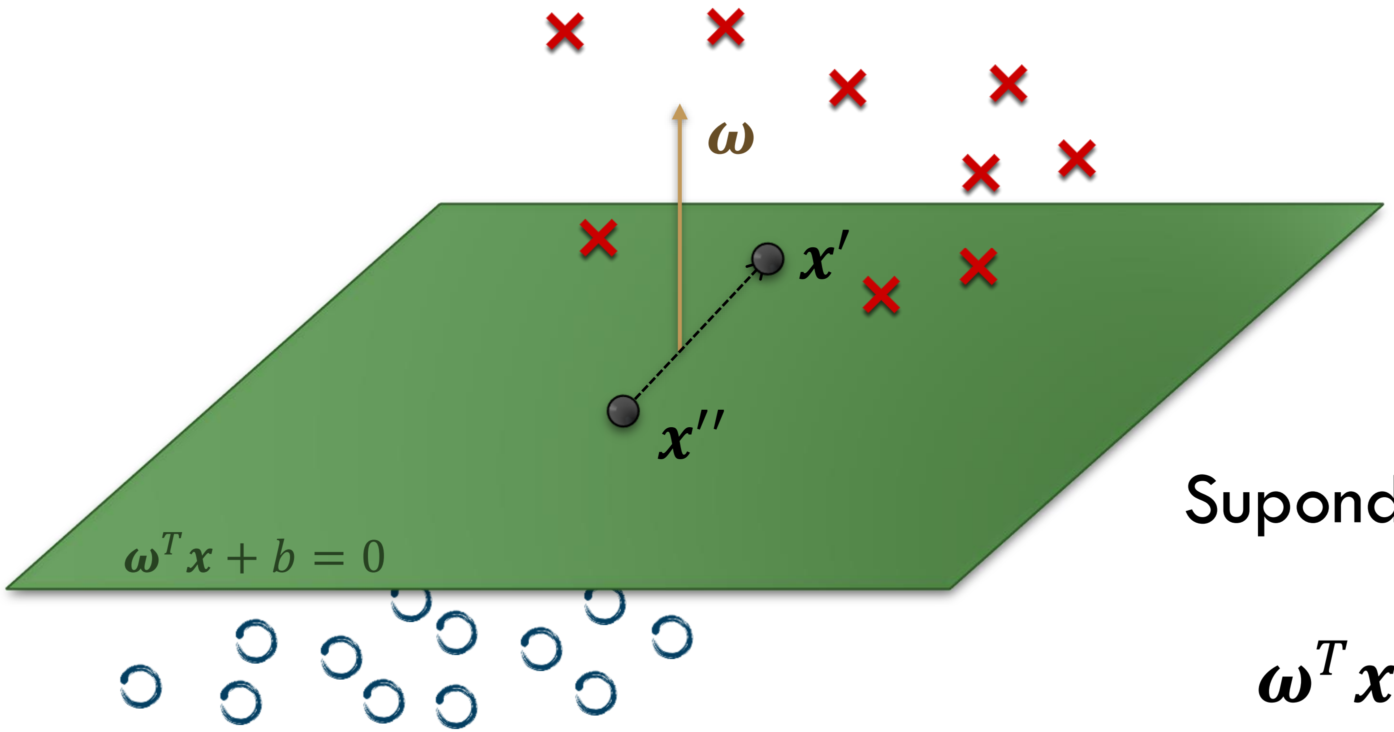
MARGEM GEOMÉTRICA γ



Como medir essa distância?

DISTÂNCIA DE UM PONTO AO PLANO

A distância entre $\boldsymbol{x}^{(i)}$ e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$



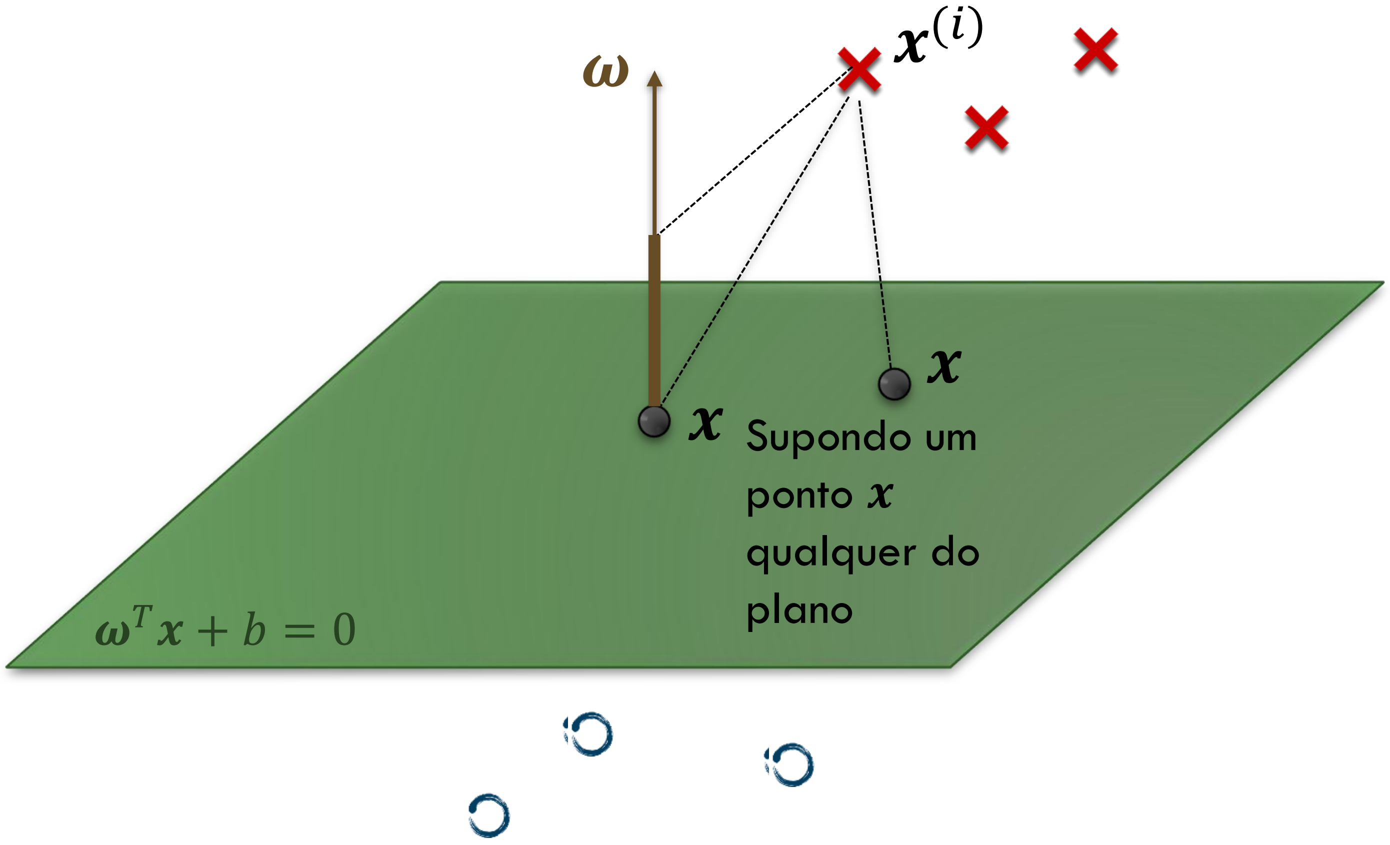
O vetor $\boldsymbol{\omega}$ é sempre perpendicular ao plano.

Supondo $\boldsymbol{x}'$ e $\boldsymbol{x}''$ pertencentes ao plano,

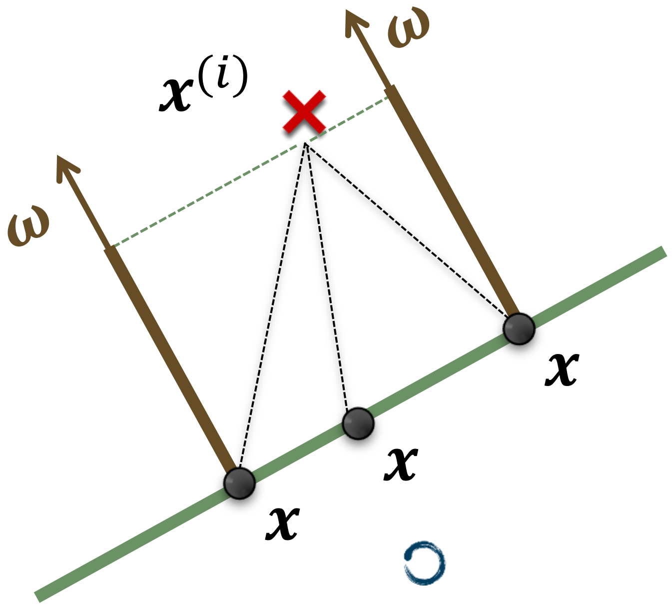
$$\boldsymbol{\omega}^T\boldsymbol{x}' + b = 0 \quad \text{e} \quad \boldsymbol{\omega}^T\boldsymbol{x}'' + b = 0$$

$$\boldsymbol{\omega}^T(\boldsymbol{x}' - \boldsymbol{x}'') = 0$$

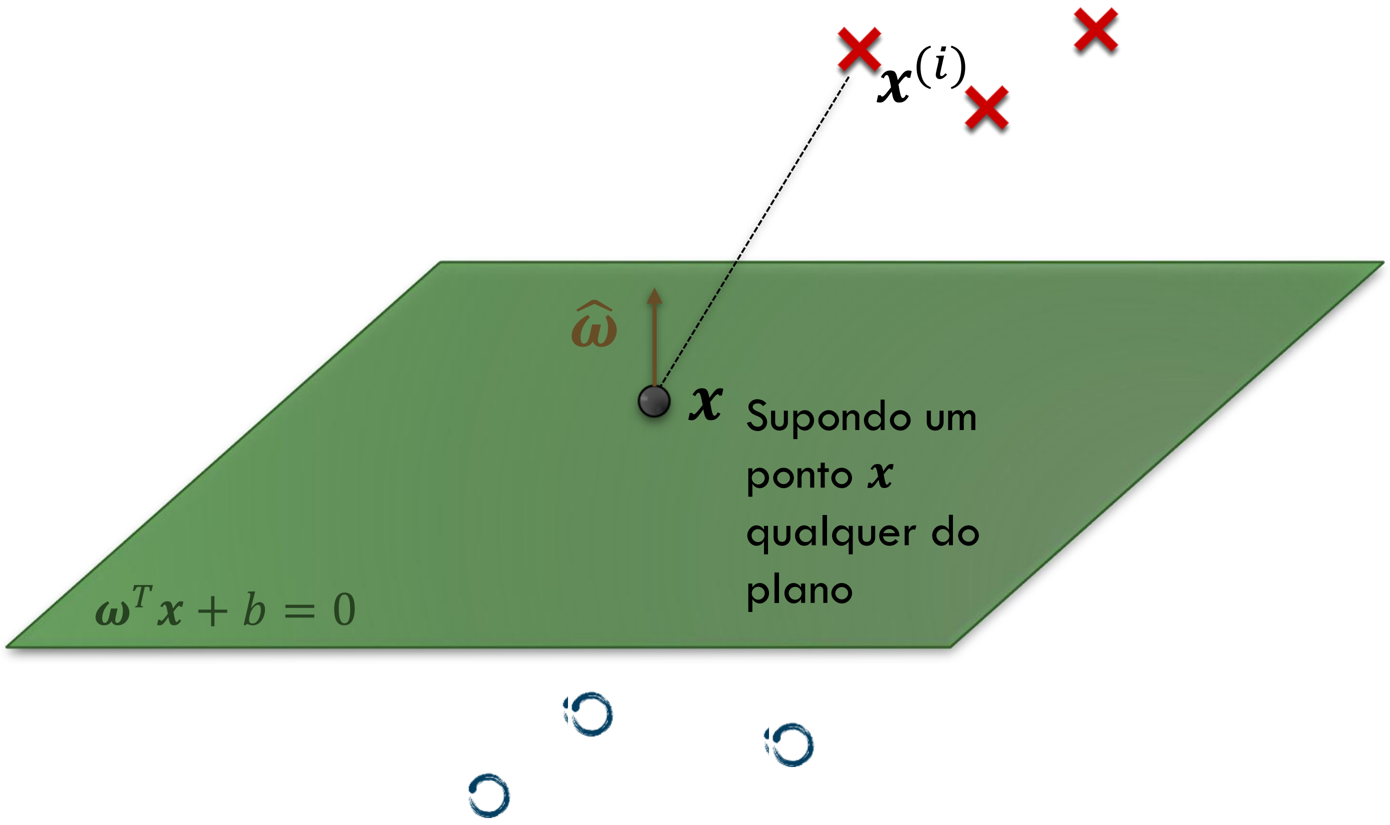
A distância entre $\boldsymbol{x}^{(i)}$ e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$



A distância euclidiana entre $\boldsymbol{x}^{(i)}$ e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$ será a projeção de $\boldsymbol{x}^{(i)} - \boldsymbol{x}$ na direção de $\boldsymbol{\omega}$



A distância entre $\boldsymbol{x}^{(i)}$ e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$



A distância euclidiana entre $\boldsymbol{x}^{(i)}$ e o plano $\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$ será a projeção de $\boldsymbol{x}^{(i)} - \boldsymbol{x}$ na direção de $\boldsymbol{\omega}$

$$\widehat{\boldsymbol{\omega}} = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|} \rightarrow \gamma^{(i)} = \left|\widehat{\boldsymbol{\omega}}^T\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}\right)\right|$$

$$\gamma^{(i)} = \frac{1}{\|\boldsymbol{\omega}\|}\left|\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\underbrace{\boxed{- \boldsymbol{\omega}^T\boldsymbol{x} - b}}_{=0}\right|$$

$$\gamma^{(i)} = \frac{1}{\|\boldsymbol{\omega}\|}y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right)$$

MARGEM GEOMÉTRICA γ

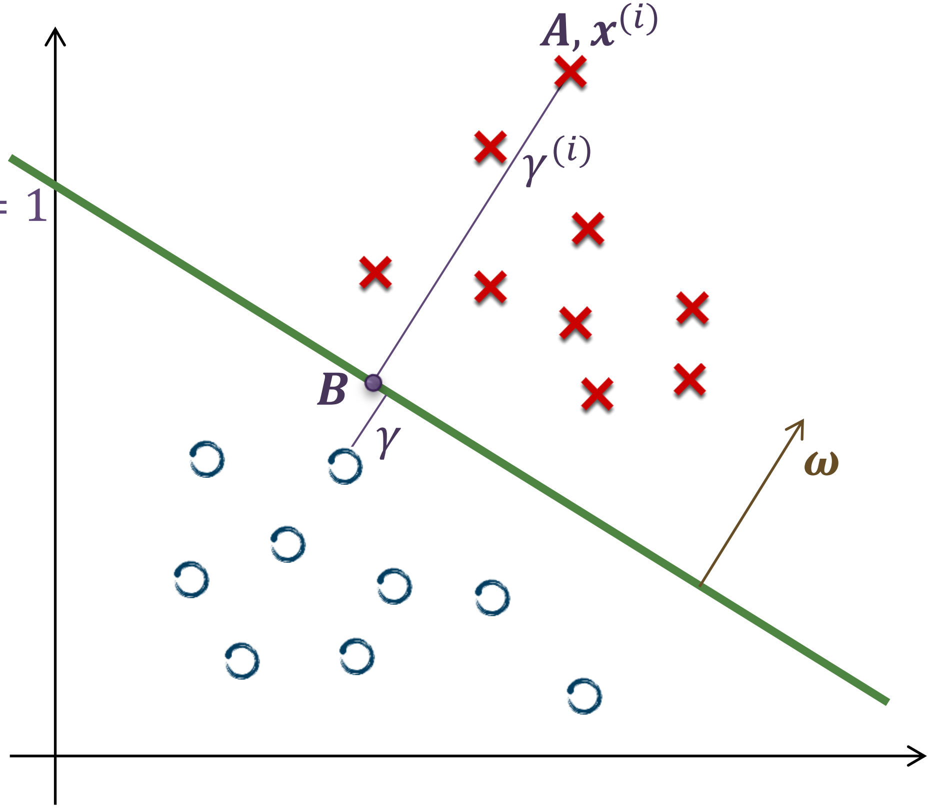
$$\gamma^{(i)} = \boxed{y^{(i)} \frac{(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b)}{\|\boldsymbol{\omega}\|}}$$

No mínimo:
$y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) = 1$

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)}$$

Define-se **margem geométrica** de $(\boldsymbol{\omega}, b)$ em relação a um conjunto de dados de treinamento como a menor das margens geométricas nos dados de treinamento individuais.

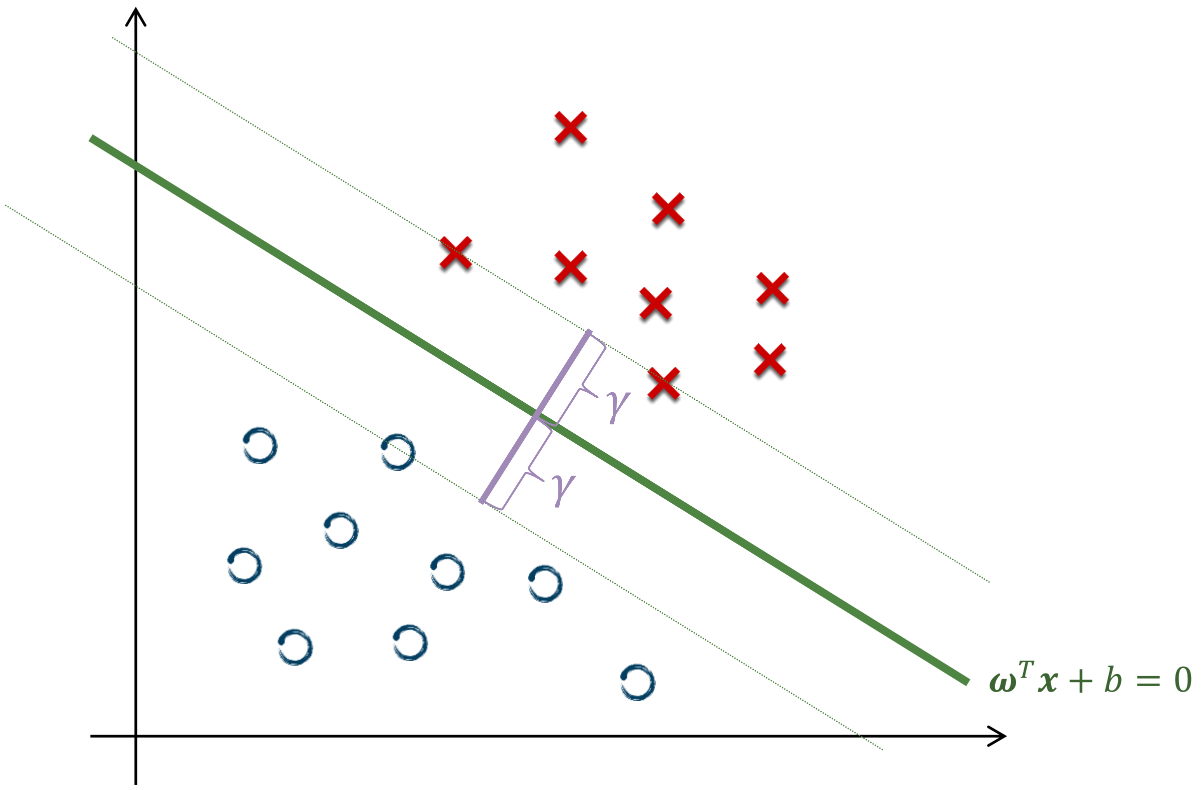
$$\gamma = \frac{1}{\|\boldsymbol{\omega}\|}$$



MARGENS

$$\gamma = \frac{1}{||\omega||}$$

Basicamente, a margem é **a terra de ninguém. Nunca haverá nenhum ponto de dados dentro da margem.**



Isso pode causar alguns problemas quando os dados são ruidosos e é por isso que o classificador de margem flexível será introduzido posteriormente.

PECE Programa de Educação Continuada
Escola Politécnica da USP

CLASSIFICADOR DE MARGEM ÓTIMA

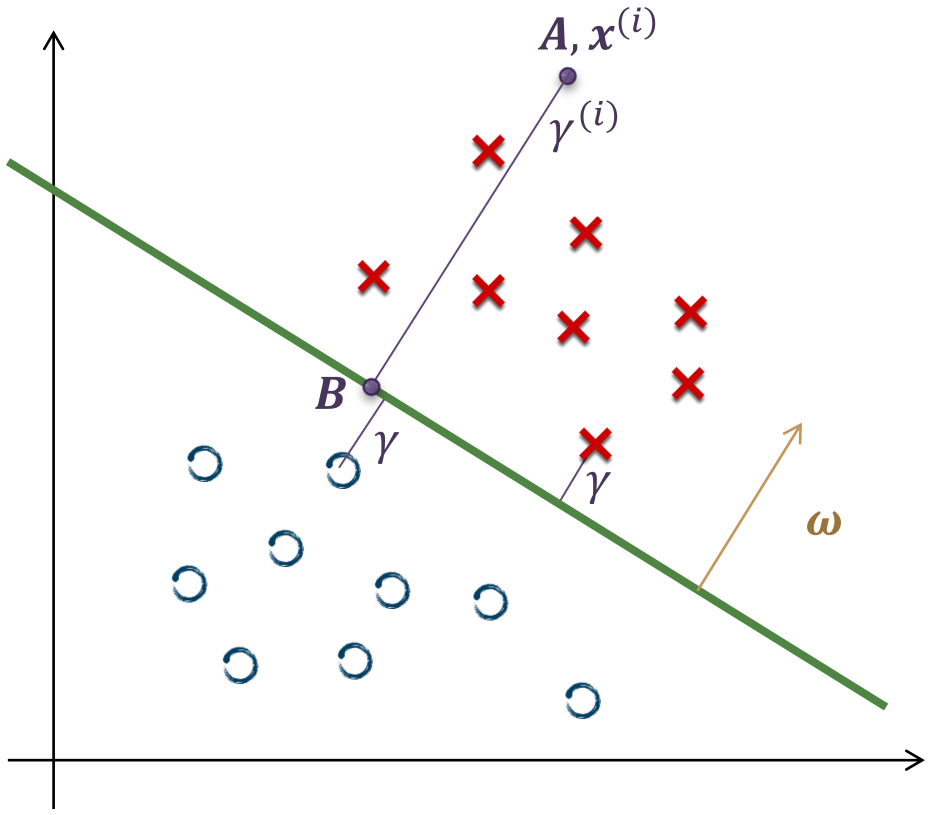
Valor ótimo sempre remete a problema de otimização

MARGEM – PROBLEMA OTIMIZAÇÃO

Maximizar $\gamma = \dfrac{1}{\|\boldsymbol{\omega}\|}$

sujeito a

$$\min_{i=1,\dots,m} y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = \mathbf{1}$$



Este não é um problema de otimização *amigável*...

Nos resta procurar um problema equivalente, mais amigável!

MARGEM ÓTIMA

$$\max \gamma = \max \frac{1}{\|\boldsymbol{\omega}\|}$$

sujeito a

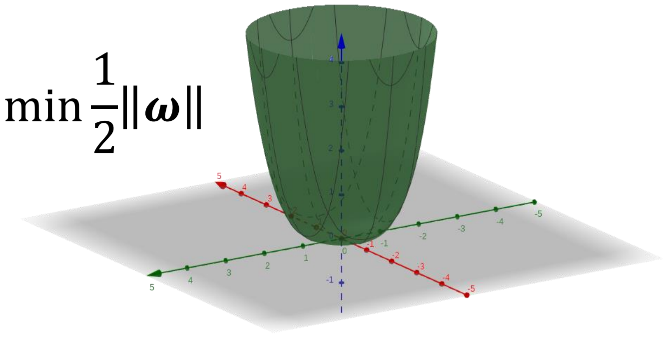
$$\min_{i=1,\ldots,m} y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = \mathbf{1}$$

$$\min \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega}$$

sujeito a,

$$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) \geq 1 \text{ para } i = 1, \ldots, m$$

$$\min \frac{1}{2} \|\boldsymbol{\omega}\|$$



Conveniência matemática...

PECE Programa de Educação Continuada
Escola Politécnica da USP

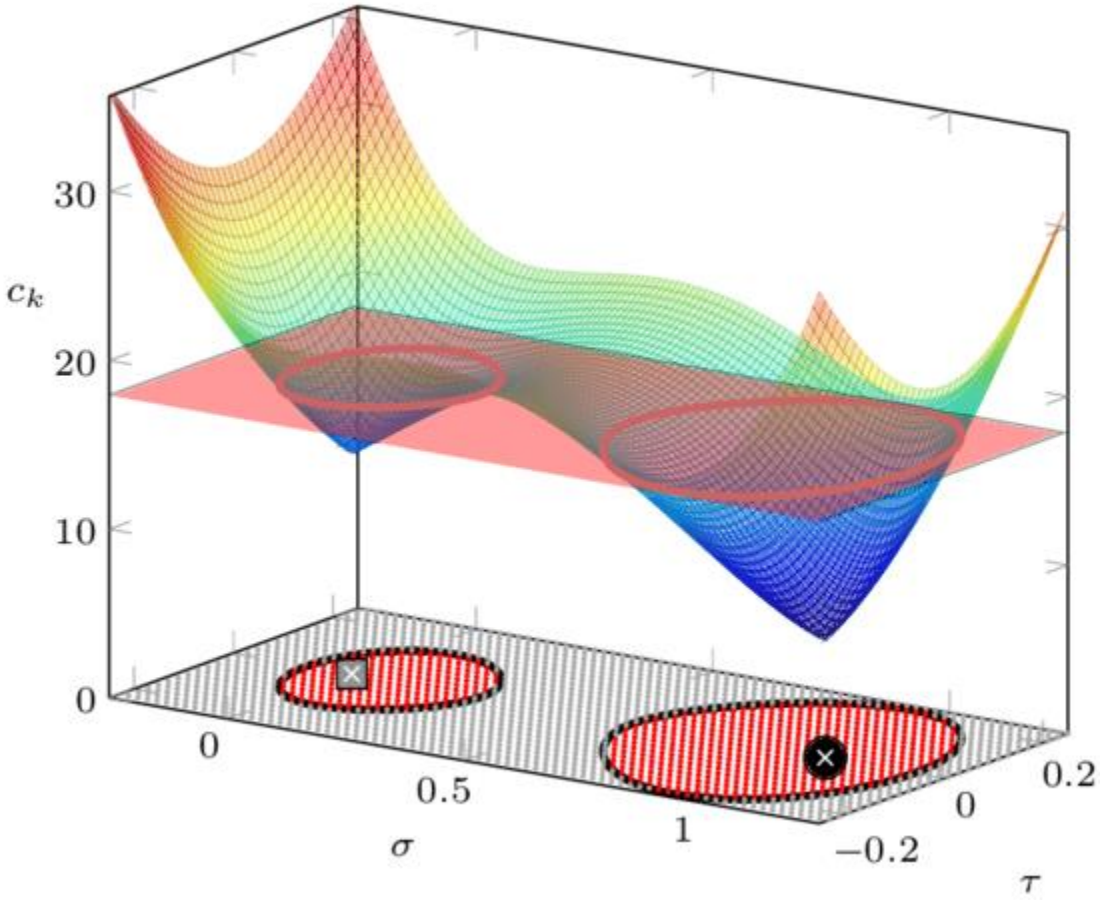
DUALIDADE DE LAGRANGE

PORQUE LAGRANGE????

Já podemos que entender que este é um problema de **otimização com restrição.**

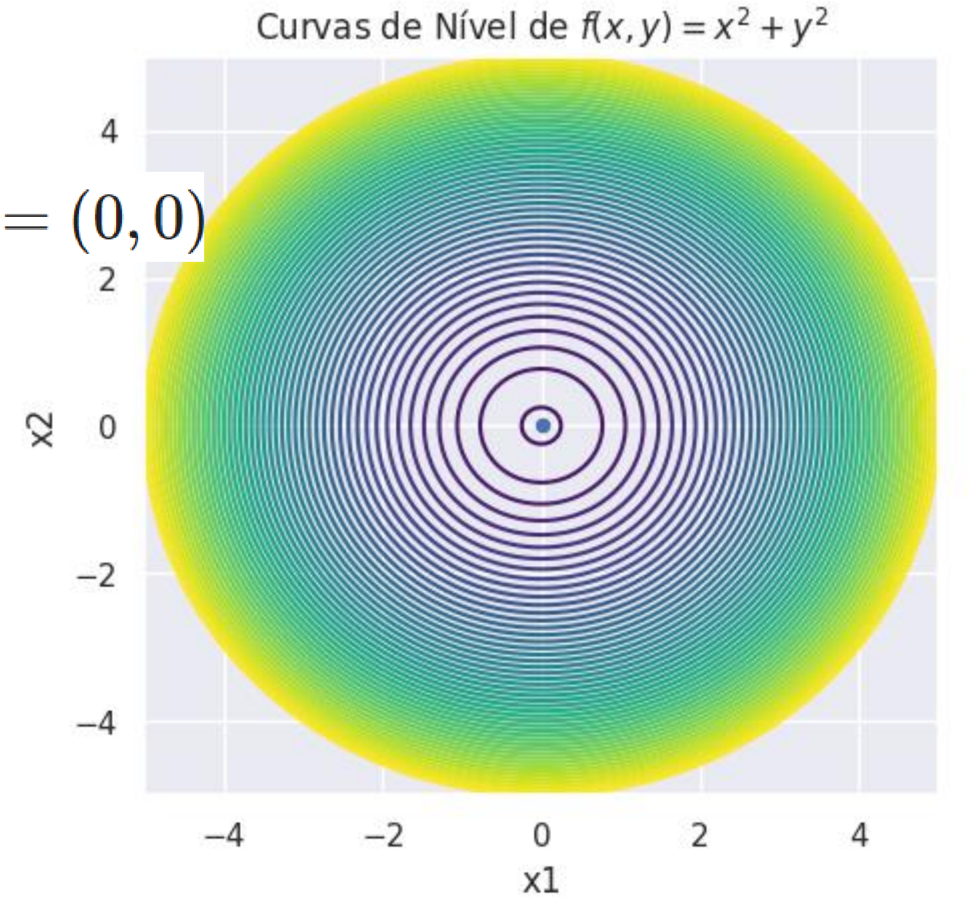
Otimização - porque, devemos encontrar a linha a partir da qual os vetores de suporte são maximamente separados

Com restrição - porque, os pontos $x^{(i)}$ devem estar afastados da margem e não dentro da margem. Usaremos multiplicadores de Lagrange para resolver esse problema.



UM PARÊNTESES SOBRE MULTIPLICADOR DE LAGRANGE

Queremos Minimizar: $f(x,y) = x^2 + y^2$ $\longrightarrow$ $(x,y) = (0,0)$



Mas, e se exigirmos que: $x + y = 1$?

UM PARÊNTESES SOBRE MULTIPLICADOR DE LAGRANGE

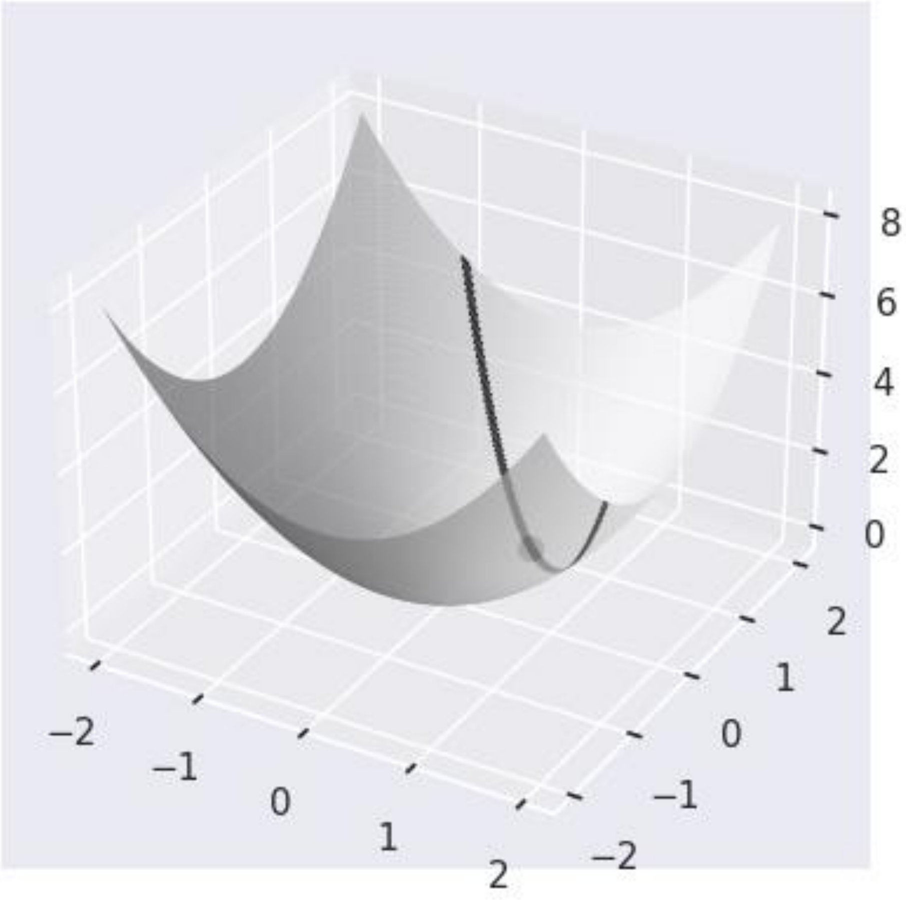
Queremos Minimizar: $f(x, y) = x^2 + y^2$

Mas, e se exigirmos que: $x + y = 1$?

$$f(x,y) = f(x, y(x)) = x^2 + (1-x)^2 = 2x^2 + 1 - 2x$$

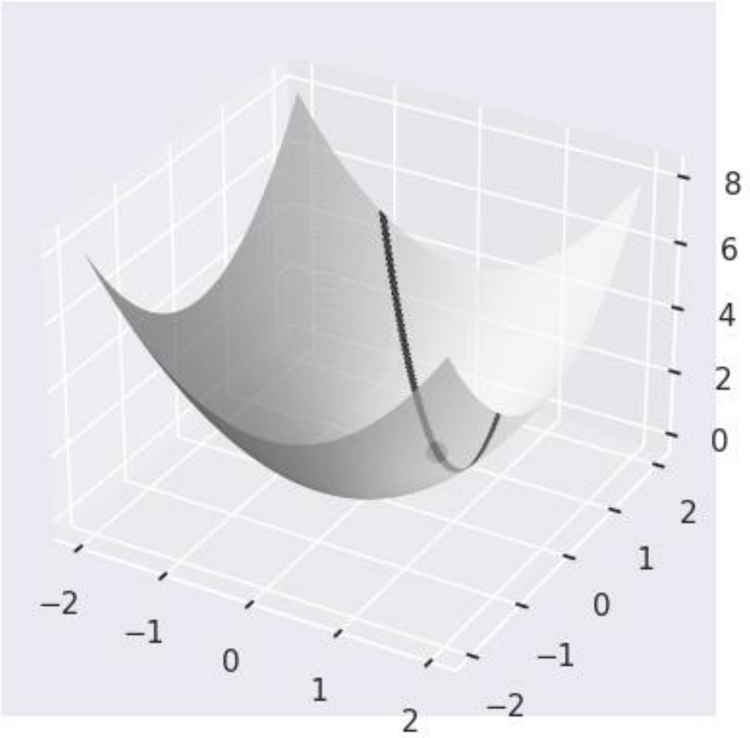
$$f'(x) = 4x - 2 = 0$$

$$(x,y) = \left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$



LAGRANGE

$$x + y - 1 = 0 = \lambda(x + y - 1)$$



$$f(x, y) = x^2 + y^2 + \lambda(x + y - 1)$$

$$\frac{\partial f}{\partial x} = 2x + \lambda = 0$$

$$\frac{\partial f}{\partial y} = 2y + \lambda = 0$$

$$\Rightarrow \quad x = y.$$

$$x + y = 1$$

$$(x, y) = \left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

$$\lambda = -1$$

λ é conhecido como um multiplicador de Lagrange. É simplesmente um truque inteligente.

Em conclusão: um multiplicador de Lagrange é uma variável que introduzimos para encontrar um extremo.

LAGRANGE

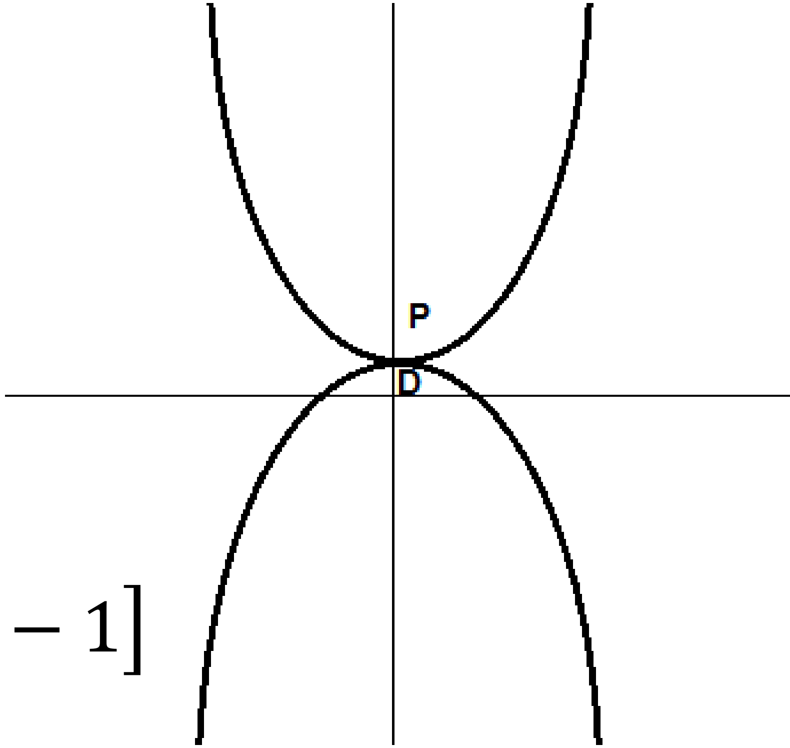
$$\min_{\omega} \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega}$$

Sujeito a

$$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) \geq 1 \quad \omega \in \mathbb{R}^d, b \in \mathbb{R}$$

Condições KKT

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} - \sum_i \alpha_i \left[y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) - 1 \right]$$



$$\min_{\omega, b} L$$

sujeito às restrições

$$\max_{\alpha \geq 0} \min_{\omega, b} L(\omega, b, \alpha)$$

DERIVAR E IGUALAR A ZERO O PROBLEMA SEM RESTRIÇÃO: $\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha})$

$$\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i\left[y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) - 1\right]$$

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = \boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i y^{(i)}\boldsymbol{x}^{(i)} = 0 \rightarrow \boxed{\boldsymbol{\omega} = \sum_{i=1}^{m} \alpha_i y^{(i)}\boldsymbol{x}^{(i)}}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \rightarrow \boxed{\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 = \boldsymbol{\alpha}^T\boldsymbol{y}}$$

SUBSTITUINDO NOSSAS DESCOBERTAS EM LAGRANGE

$$\boldsymbol{\omega} = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b \right) - 1 \right]$$

$$\boldsymbol{\omega}^T \boldsymbol{\omega} = \sum_{j=1}^{m} \alpha_j y^{(j)} \boldsymbol{x}^{(j)^T} \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(j)^T} \boldsymbol{x}^{(i)}$$

$$= \sum_{i=1}^{m} \alpha_i y^{(i)} \left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b \right) - \alpha_i$$

$$= \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + \cancelto{0}{\sum_{i=1}^{m} \alpha_i y^{(i)} b} - \sum_{i=1}^{m} \alpha_i$$

$$= \boldsymbol{\omega}^T \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} - \sum_{i=1}^{m} \alpha_i = \boldsymbol{\omega}^T \boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i$$

SUBSTITUINDO NOSSAS DESCOBERTAS EM LAGRANGE

$$\boldsymbol{\omega} = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) - 1\right]$$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)}\, \alpha_i \alpha_j {\boldsymbol{x}^{(j)}}^T \boldsymbol{x}^{(i)}$$

LAGRANGE E SEU PROBLEMA DUAL

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)^T} \boldsymbol{x}^{(j)}$$

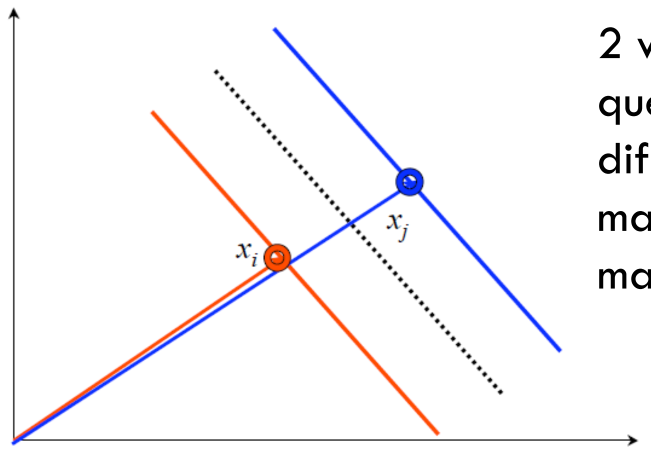
sujeito a

$$\alpha_i \geq 0, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

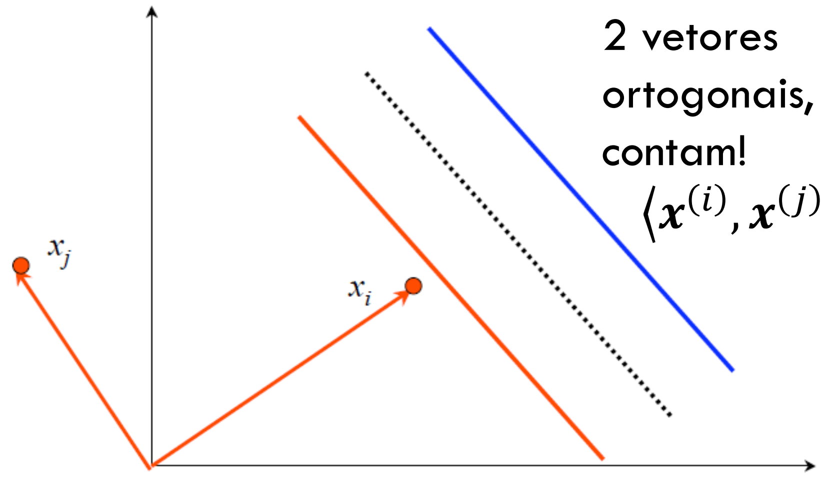
INTUIÇÃO....

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} {\boldsymbol{x}^{(i)}}^T \boldsymbol{x}^{(j)}$$



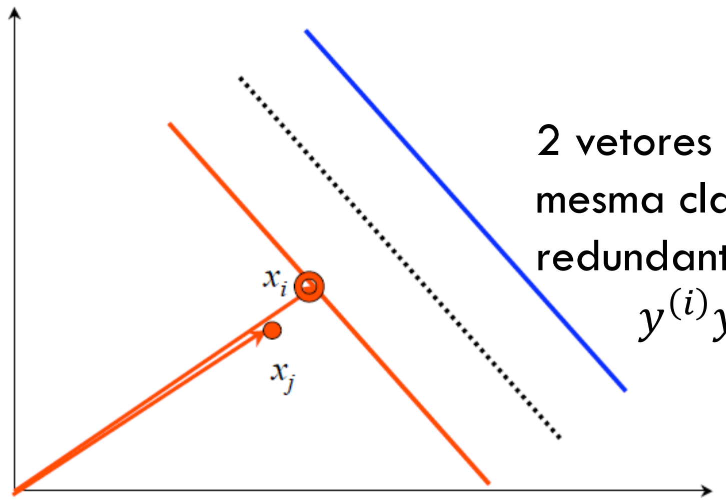
2 vetores muito similares que predizem classes diferentes tendem a maximizar a largura da margem.

$$y^{(i)} y^{(j)} < 0$$



2 vetores ortogonais, não contam!

$$\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \rangle = 0$$



2 vetores similares, da mesma classe, são redundantes!

$$y^{(i)} y^{(j)} > 0$$

LAGRANGE E SEU PROBLEMA DUAL

$$\min_{\boldsymbol{\alpha}}[-L(\boldsymbol{\alpha})] = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} - \sum_{i=1}^{m}\alpha_i$$

sujeito a

$$\alpha_i \geq 0, i = 1, \dots, m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

PECE Programa de Educação Continuada
Escola Politécnica da USP

O QUE O PROGRAMA FAZ

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \underbrace{\begin{bmatrix} y^{(1)}y^{(1)}\boldsymbol{x}^{(1)}\cdot\boldsymbol{x}^{(1)} & y^{(1)}y^{(2)}\boldsymbol{x}^{(1)}\cdot\boldsymbol{x}^{(2)} & \cdots & y^{(1)}y^{(m)}\boldsymbol{x}^{(1)}\cdot\boldsymbol{x}^{(m)} \\ y^{(2)}y^{(1)}\boldsymbol{x}^{(2)}\cdot\boldsymbol{x}^{(1)} & y^{(2)}y^{(2)}\boldsymbol{x}^{(2)}\cdot\boldsymbol{x}^{(2)} & \cdots & y^{(2)}y^{(m)}\boldsymbol{x}^{(2)}\cdot\boldsymbol{x}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ y^{(m)}y^{(1)}\boldsymbol{x}^{(m)}\cdot\boldsymbol{x}^{(1)} & y^{(m)}y^{(2)}\boldsymbol{x}^{(m)}\cdot\boldsymbol{x}^{(2)} & \cdots & y^{(m)}y^{(m)}\boldsymbol{x}^{(m)}\cdot\boldsymbol{x}^{(m)} \end{bmatrix}}_{\text{Coeficientes quadráticos}} \boldsymbol{\alpha} + \underbrace{(-1 \quad \cdots \quad -1)\boldsymbol{\alpha}}_{\text{Linear}}$$

sujeito a

$$\underbrace{\boldsymbol{y}^T\boldsymbol{\alpha} = 0}_{\text{Restrição Linear}}$$

$$\underbrace{\boldsymbol{0} \leq \boldsymbol{\alpha} \leq \infty,}_{\text{Limites inferior e superior}} i = 1, \ldots, m$$

O que o problema de otimização devolve???? $\boldsymbol{\alpha}$

O que a biblioteca em python devolve???? $\boldsymbol{b}, \boldsymbol{\omega}$

A SOLUÇÃO DE LAGRANGE NOS DEVOLVEU α

PECE Programa de Educação Continuada
Escola Politécnica da USP

$$\boldsymbol{\alpha} = \alpha_1, \alpha_2, \dots, \alpha_m$$

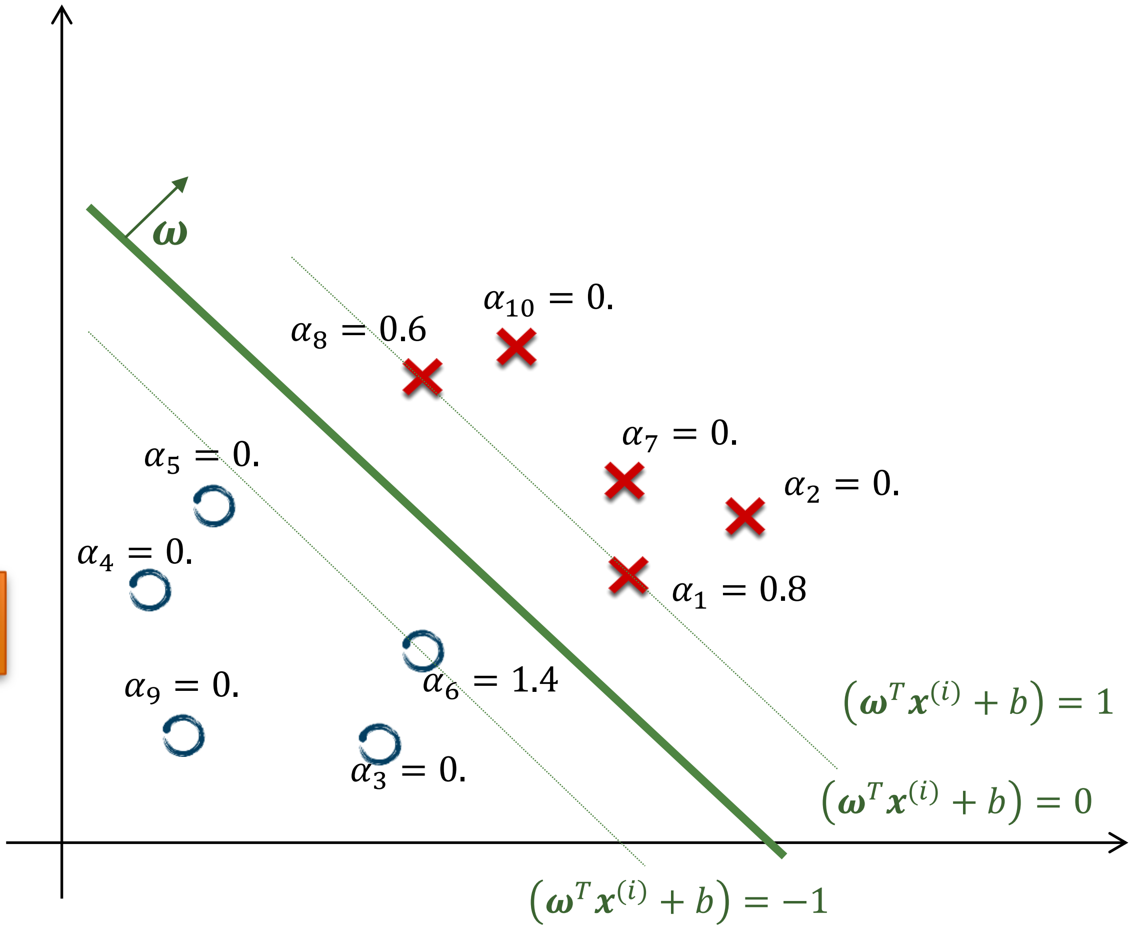
$$\boldsymbol{\omega} = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

Os vetores de suporte são os únicos que exercem influência na construção do hiperplano de máxima margem.

Condições KKT: $\alpha_i\left[y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) - 1\right] = 0$

Os dados de entrada com a margem igual a 1 são chamados de vetores-suporte, sendo justamente aqueles com $\alpha_i \neq 0$.

$\alpha_i > 0$, então i é um vetor de suporte

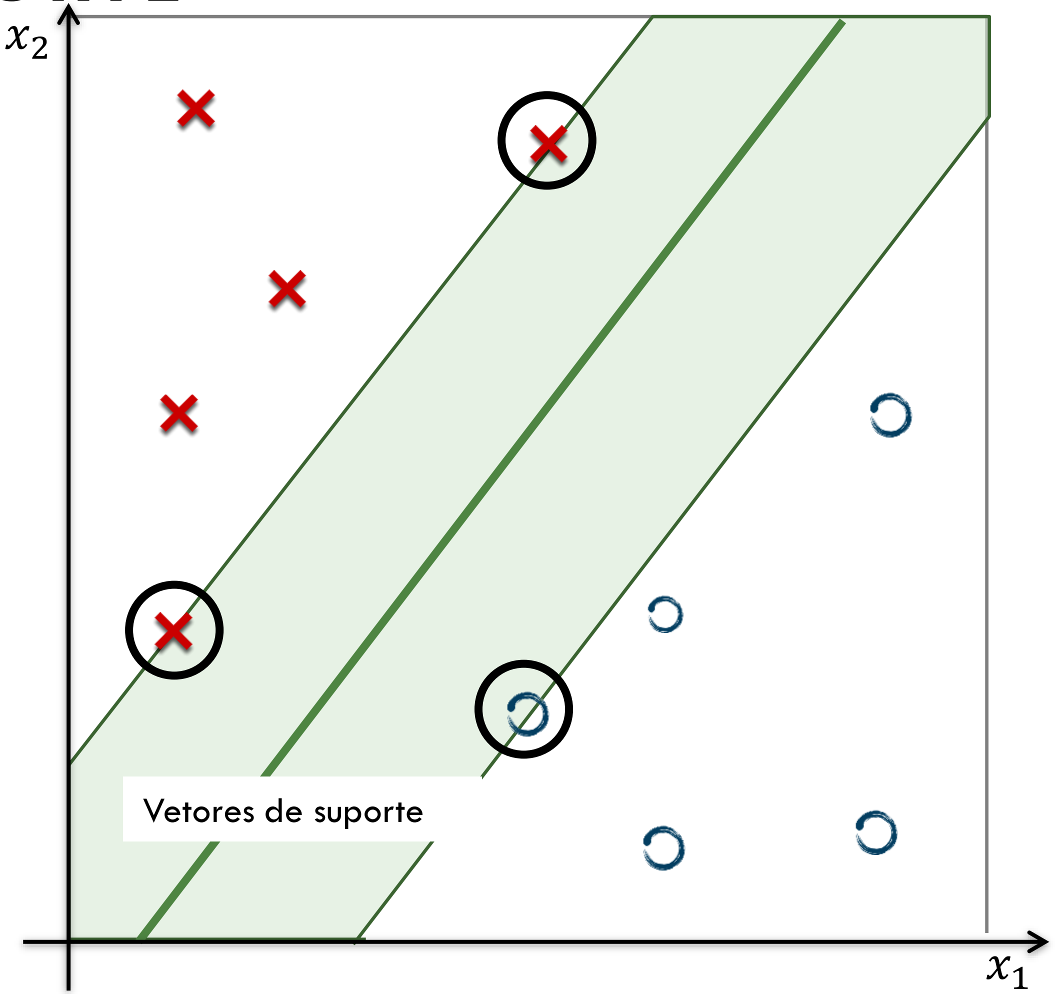


MAIS SOBRE VETORES DE SUPORTE

Os vetores que atingem a menor distância são chamados de **vetores de suporte.** São pontos de dados de treinamento que estão mais próximos do hiperplano e influenciam a posição e a orientação do hiperplano. Usando esses vetores de suporte, maximizamos a margem do classificador.

Os vetores de suporte são os exemplos mais difíceis de classificar. São os elementos críticos do conjunto de treinamento

Somente um subconjunto muito pequeno de amostras de treinamento (vetores de suporte) pode especificar completamente a função de decisão. Os pontos interiores não contribuem em nada para solução!



ENTÃO...

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

Em qualquer vetor de suporte i,

$$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

Usamos essa equação para achar o valor de b

RESUMO:

PECE Programa de Educação Continuada
Escola Politécnica da USP

FORMULAÇÃO DO PROBLEMA: O objetivo é encontrar um limite de decisão (um hiperplano) que maximize a margem entre as duas classes.

$D = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$, onde $\boldsymbol{x}^{(i)}$ é o vetor de características i, e $y^{(i)}$ sua classe ($+1$ ou -1)

DEFINIÇÃO O HIPERPLANO: Um hiperplano é definido como $\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b = 0$, onde $\boldsymbol{\omega}$ é o vetor de pesos normal ao hiperplano e b é o viés.

$y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) = 1$ se $\boldsymbol{x}^{(i)}$ é um vetor de suporte
$y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) > 1$ se $\boldsymbol{x}^{(i)}$ não é um vetor de suporte

MAXIMIZAR A MARGEM: A margem pode ser calculada como a distância perpendicular entre os vetores de suporte e o hiperplano. Para maximizar a margem, precisamos minimizar:

$$\min_{\boldsymbol{\omega}} \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega}$$

Sujeito a

$y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) \geq 1$ com $\omega \in \mathbb{R}^d, b \in \mathbb{R}$

Formulação dual: Resolva o problema de programação quadrática convexa usando técnicas de otimização, como o algoritmo Sequential Minimal Optimization (SMO), ou gradiente ascendente.

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)^T} \boldsymbol{x}^{(j)}$$

sujeito a $\alpha_i \geq 0, \ \sum \alpha_i y^{(i)} = 0$

Obtenção dos pesos e viés $\boldsymbol{\omega}$ e $\boldsymbol{b}$ a partir das equações:

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \qquad y^{(i)}(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b) = 1$$

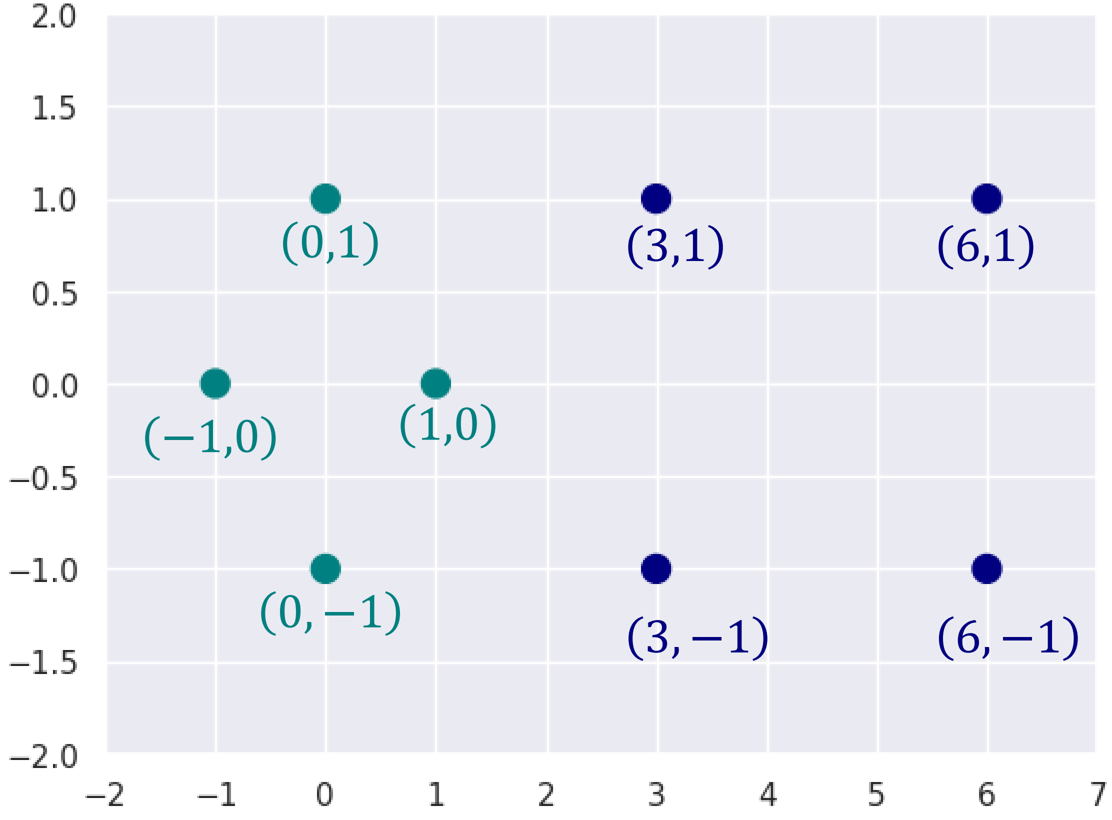
Fazer previsões: Para um novo ponto de dados $\boldsymbol{x}$, o rótulo de classe previsto ŷ pode ser calculado usando:

$$h(\boldsymbol{x}) = \text{sign}(\boldsymbol{\omega}^T \boldsymbol{x} + b)$$

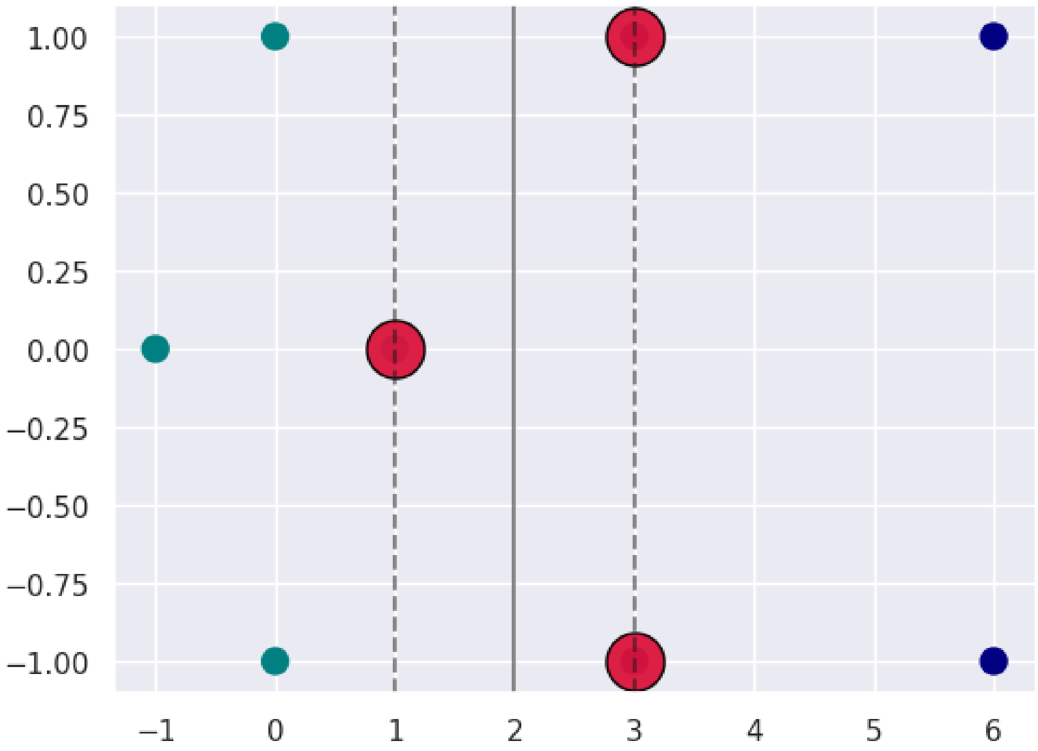
EXEMPLO NOTEBOOK

$$x_+ = \left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 6 \\ -1 \end{pmatrix} \right\}$$

$$x_- = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right\}$$



A SOLUÇÃO É SIMPLES....



Vamos passo a passo seguir nossa metodologia, para chegarmos à essa conclusão matematicamente...

```
Coeficientes do modelo
w1, w2 =[[ 1.00048166e+00 -1.66533454e-16]]
b = [-2.00112388]
```

PRIMEIRAMENTE, SEM RESOLVER A OTIMIZAÇÃO

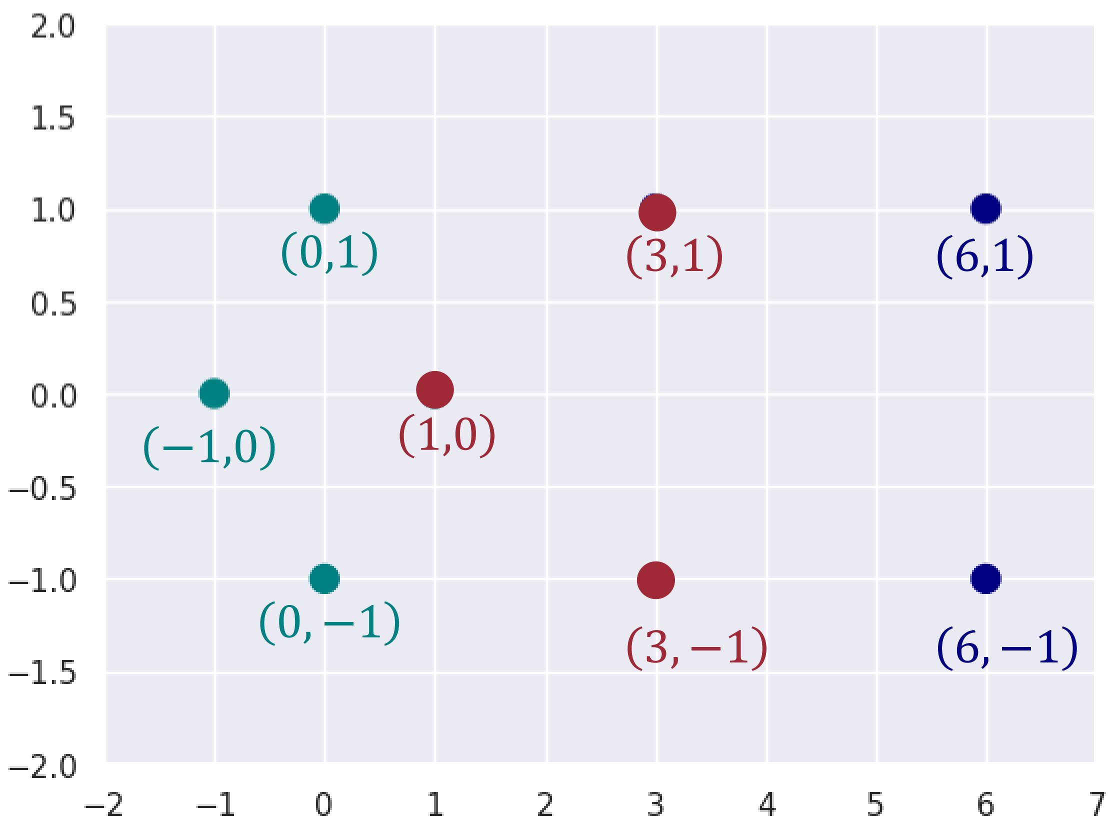
Por inspeção, verifica-se que os vetores de suporte são 3: $(1{,}0)$ para $y = -1$; $(3{,}1)$, para $y = +1$; e $(3,-1)$, para $y = +1$. Dessa forma, pode-se escrever os vetores de suporte $sv_i = 1, \cdots, 3$ como,

$$sv_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad sv_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad sv_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

Se:

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \qquad y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

$$\omega^T x^{(j)} + b = \sum_{i=1}^{n_{sv}} \alpha_i y^{(i)} \langle x^{(i)}, x^{(j)} \rangle + b = y^{(i)}$$



$$\sum_{i=1}^{n_{sv}} \alpha_i y^{(i)} \langle x^{(i)}, x^{(j)} \rangle + b = y^{(i)}$$

$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(1)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(1)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(1)} + b = y^{(1)}$$
$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(2)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(2)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(2)} + b = y^{(2)}$$
$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(3)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(3)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(3)} + b = y^{(3)}$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = -\alpha_1 + \alpha_2 + \alpha_3 = 0$$

$$sv_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad sv_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad sv_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

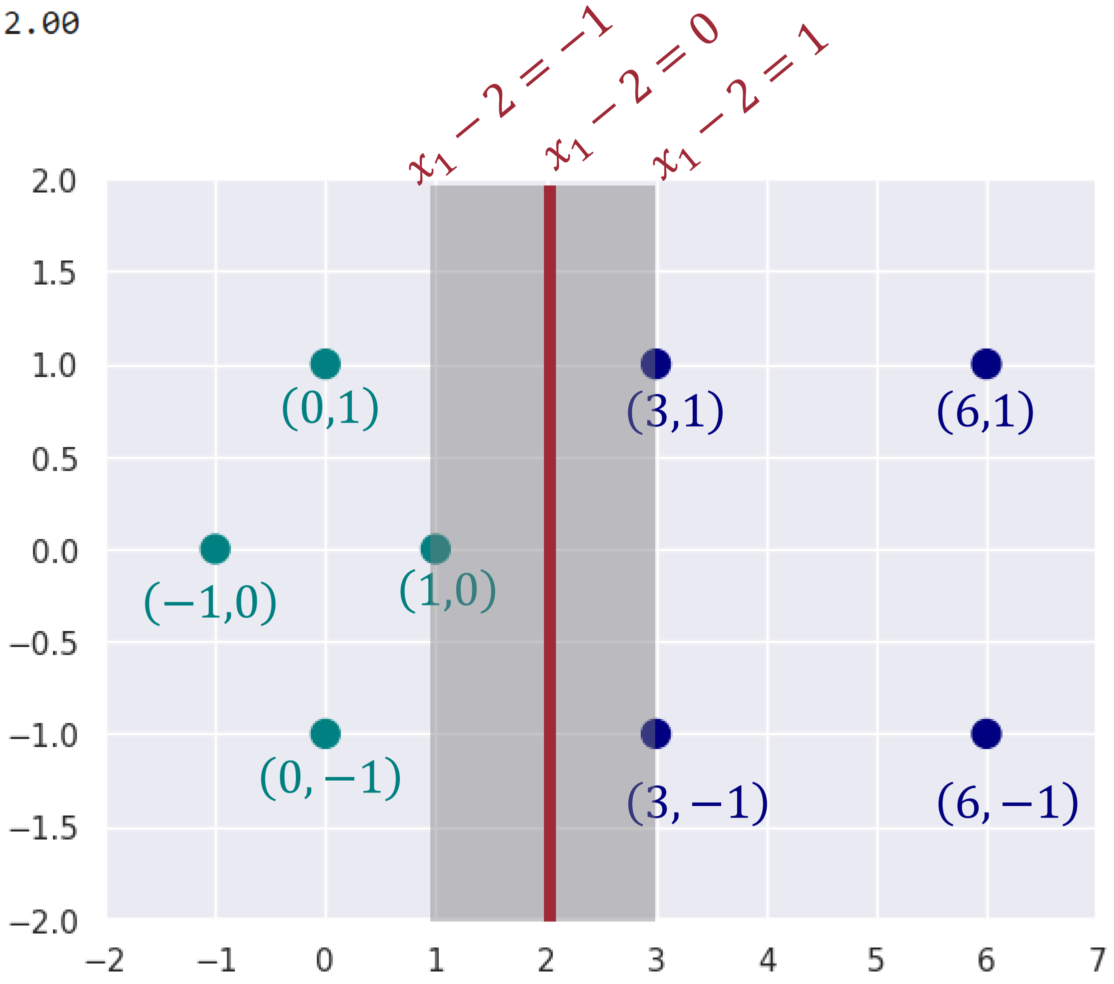
$\langle x^{(1)}, x^{(1)} \rangle = 1$, $\langle x^{(2)}, x^{(2)} \rangle = 10$, $\langle x^{(3)}, x^{(3)} \rangle = 10$, $\langle x^{(1)}, x^{(2)} \rangle = 3$, $\langle x^{(2)}, x^{(3)} \rangle = 8$ e $\langle x^{(1)}, x^{(3)} \rangle = 3$,

$$-1\alpha_1 + 3\alpha_2 + 3\alpha_3 + b = -1$$
$$-3\alpha_1 + 10\alpha_2 + 8\alpha_3 + b = +1$$
$$-3\alpha_1 + 8\alpha_2 + 10\alpha_3 + b = +1$$
$$-\alpha_1 + \alpha_2 + \alpha_3 + 0b = 0$$

```
[[-1.  3.  3.  1.]
 [-3. 10.  8.  1.]
 [-3.  8. 10.  1.]
 [-1.  1.  1.  0.]]
alpha 1 = 0.50, alpha 2 = 0.25, alpha 3 = 0.25
b = -2.00
```

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

```
w 1 = 1.00, w 2 = 0.00
```

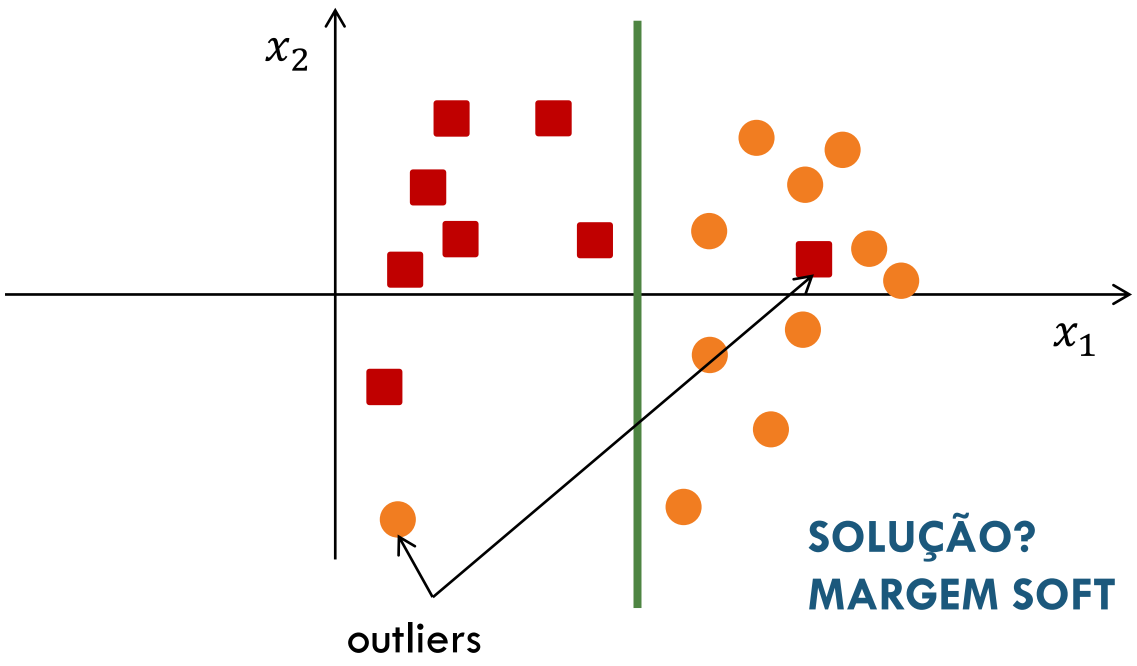


PECE Programa de Educação Continuada
Escola Politécnica da USP

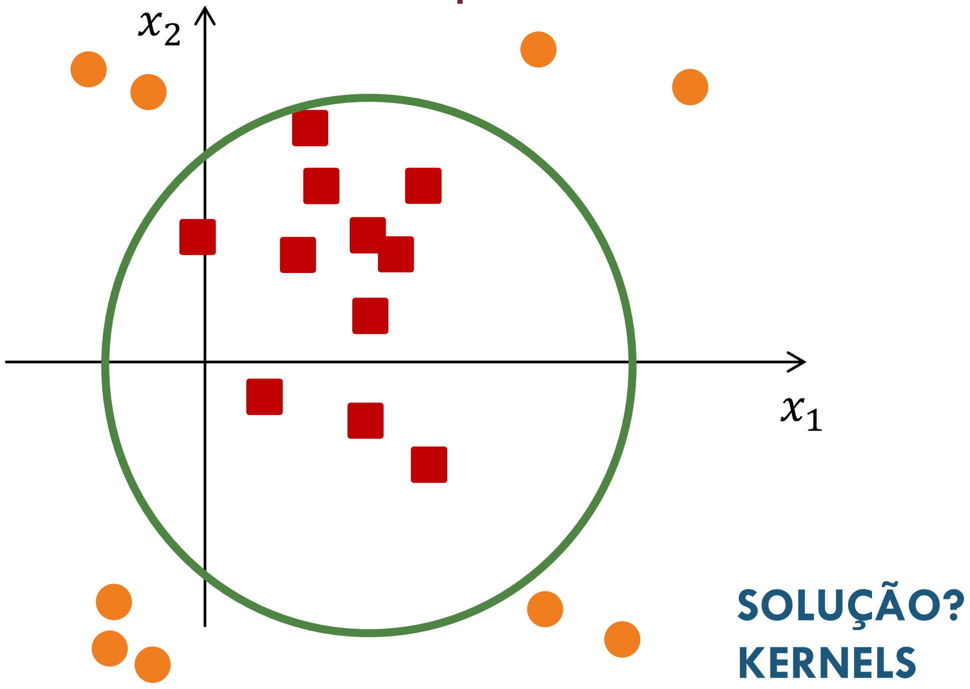
MARGEM SOFT

DOIS TIPOS DE DADOS **NÃO SEPARÁVEIS**

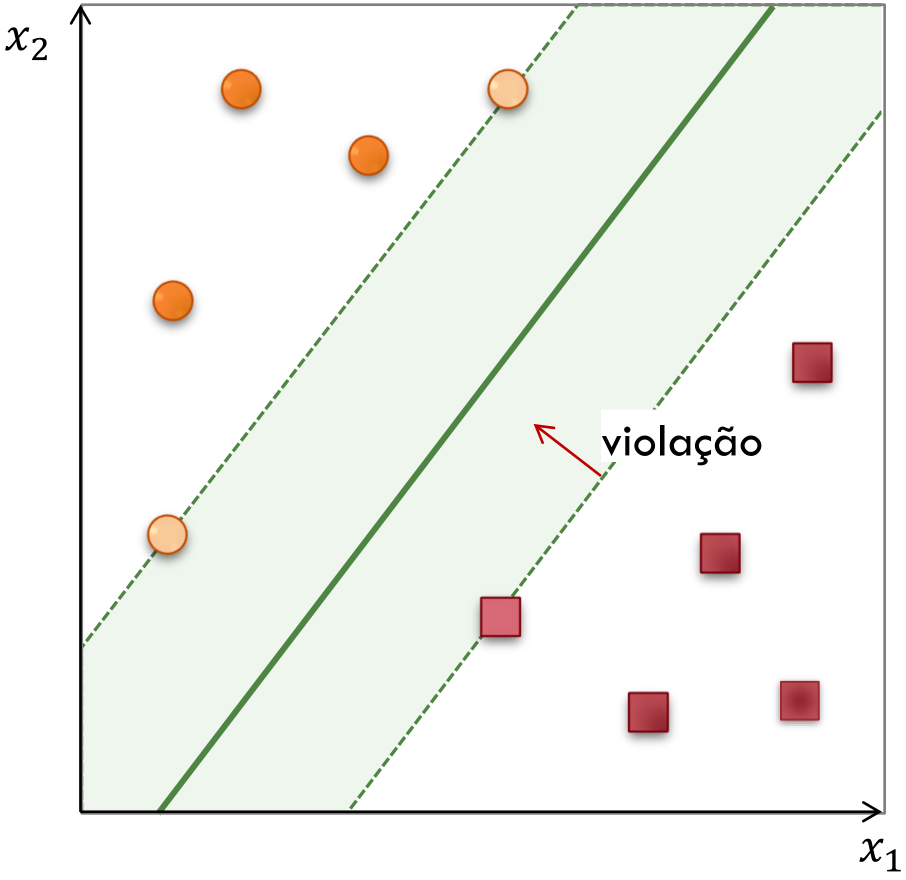
Levemente não separáveis...



Seriamente não separáveis



VIOLAÇÃO DE MARGEM



$$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)}$$

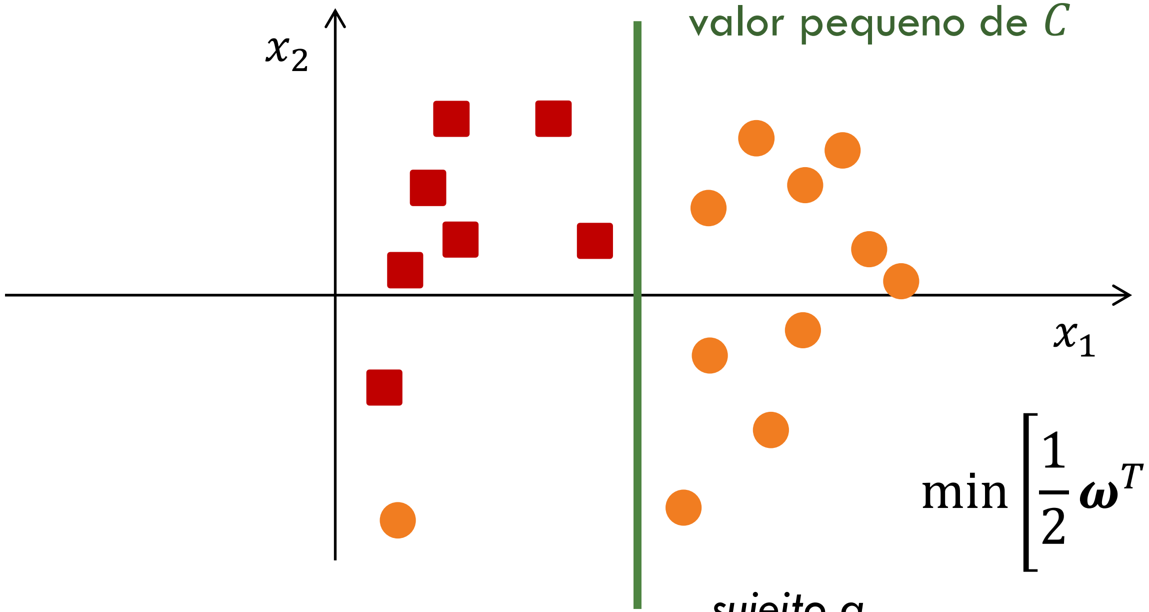
Folga ou relaxamento que estou dando à minha restrição

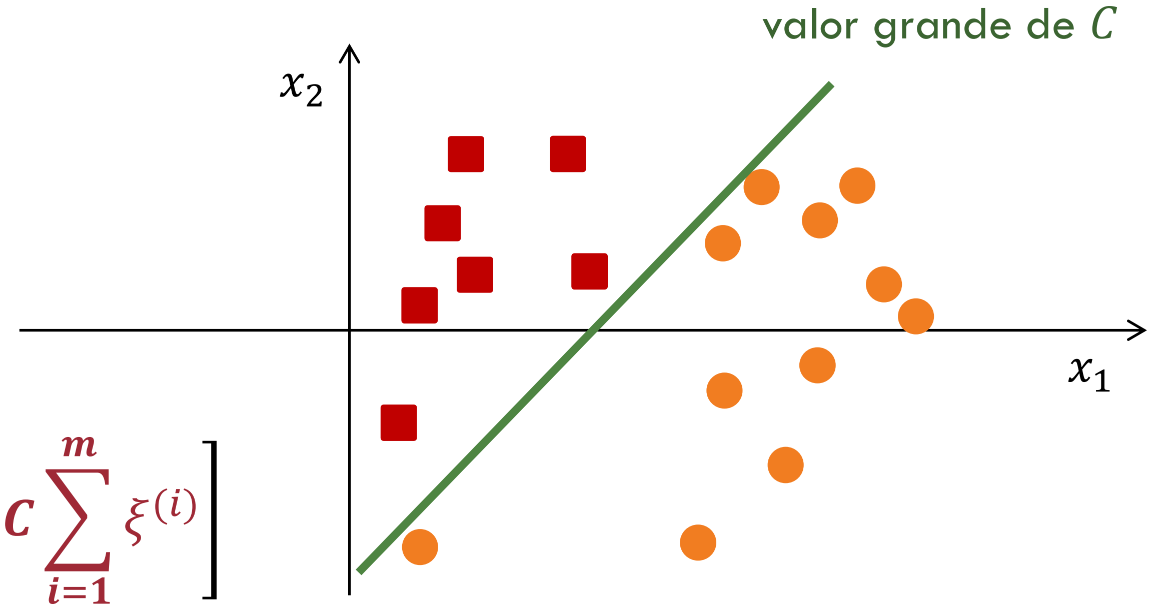
$$\xi^{(i)} \geq 0, i = 1, \dots, m$$

Violação total:

$$\sum_{i=1}^{m} \xi^{(i)}$$

MARGEM SOFT (REGULARIZAÇÃO L1)





$$\min\left[\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \boldsymbol{C}\sum_{i=1}^{m}\xi^{(i)}\right]$$

sujeito a,

$$y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)}$$

$$\xi^{(i)} \geq 0, i = 1, \dots, m$$

REGULARIZAÇÃO

$$\min\left[\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \boldsymbol{C}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \xi^{(i)}\right]$$

Isso significa que encontramos uma linha de separação que penaliza pontos no "lado errado“. I.é,

- Agora é permitido que exemplos tenham margem funcional menor que1
- Se o exemplo tiver margem $1-\xi^{(i)}$, pagamos o custo do objetivo sendo aumentado por $C\xi^{(i)}$

LAGRANGE E SEU PROBLEMA DUAL

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^{m}\xi^{(i)} - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi^{(i)}\right] - \sum_{i=1}^{m}\beta_i\xi^{(i)}$$

Minimizar com respeito a $\boldsymbol{\omega}, b, \boldsymbol{\xi}$ e maximizar com respeito a cada $\alpha_i \geq 0$ e $\beta_i \geq 0$

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = \boldsymbol{\omega} - \sum_{i=1}^{m}\alpha_i y^{(i)}\boldsymbol{x}^{(i)} = 0 \rightarrow \boxed{\boldsymbol{\omega} = \sum_{i=1}^{m}\alpha_i y^{(i)}\boldsymbol{x}^{(i)}}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m}\alpha_i y^{(i)} = 0 \rightarrow \sum_{i=1}^{m}\alpha_i y^{(i)} = \boxed{0 = \boldsymbol{\alpha}^T\boldsymbol{y}}$$

$$\frac{\partial L}{\partial \xi^{(i)}} = \boxed{C - \alpha_i - \beta_i = 0}$$

LAGRANGE E SEU PROBLEMA DUAL

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^{m}\xi^{(i)} - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(\boldsymbol{\omega}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi^{(i)}\right] - \sum_{i=1}^{m}\beta_i\xi^{(i)}$$

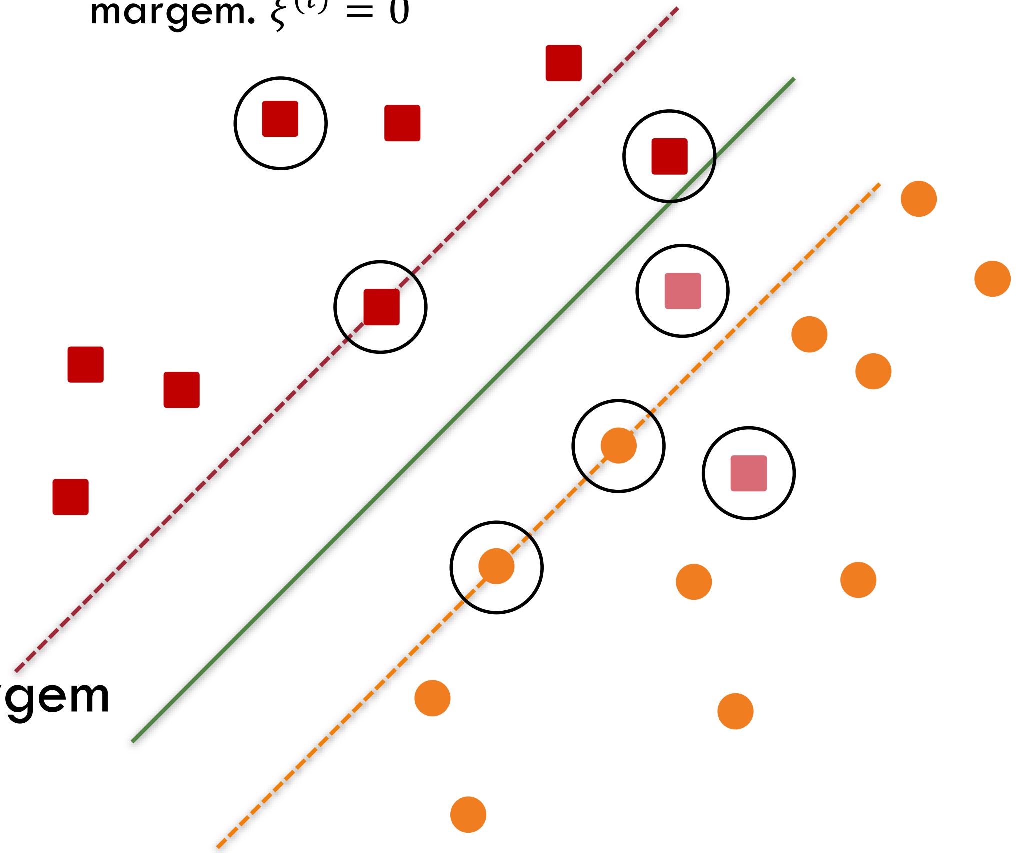
$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)}$$

sujeito a

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \quad , \sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

TIPOS DE VETORES DE SUPORTE

$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) > 1$: O ponto está além da margem. $\xi^{(i)} = 0$



Vetores de suporte **de margem** $(0 < \alpha_i < C)$:

$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = 1$: ponto está na margem

$\xi^{(i)} = 0$ Nenhuma contribuição para a perda, como no caso de margem rígida

Vetor de suporte **fora da margem** $(\alpha_i = C)$:

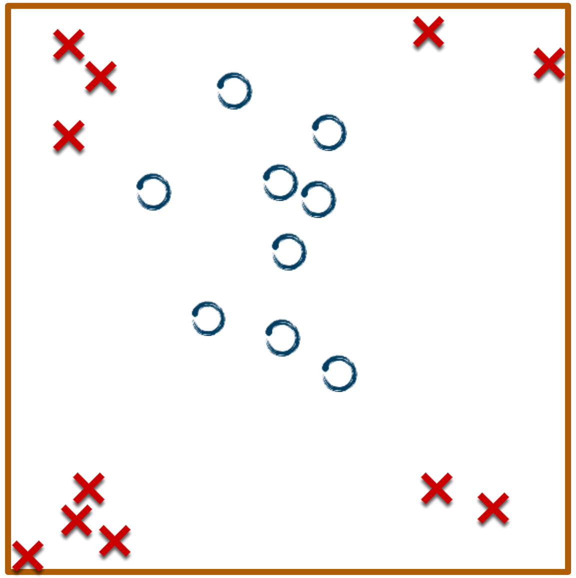
$y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) < 1$: ponto viola a restrição de margem

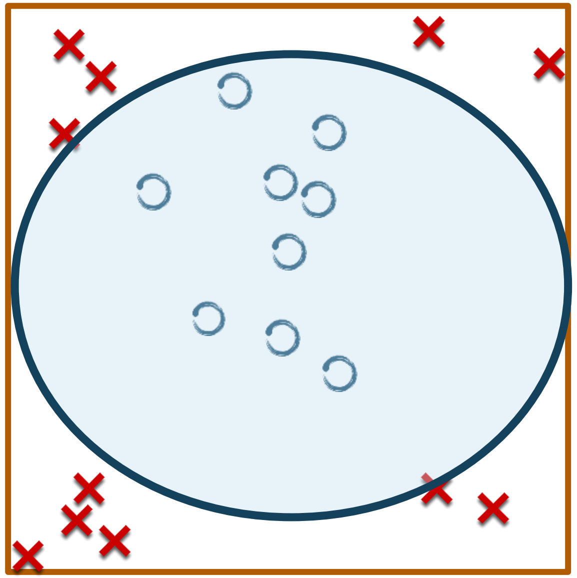
$\xi^{(i)} > 0$

PECE Programa de Educação Continuada
Escola Politécnica da USP

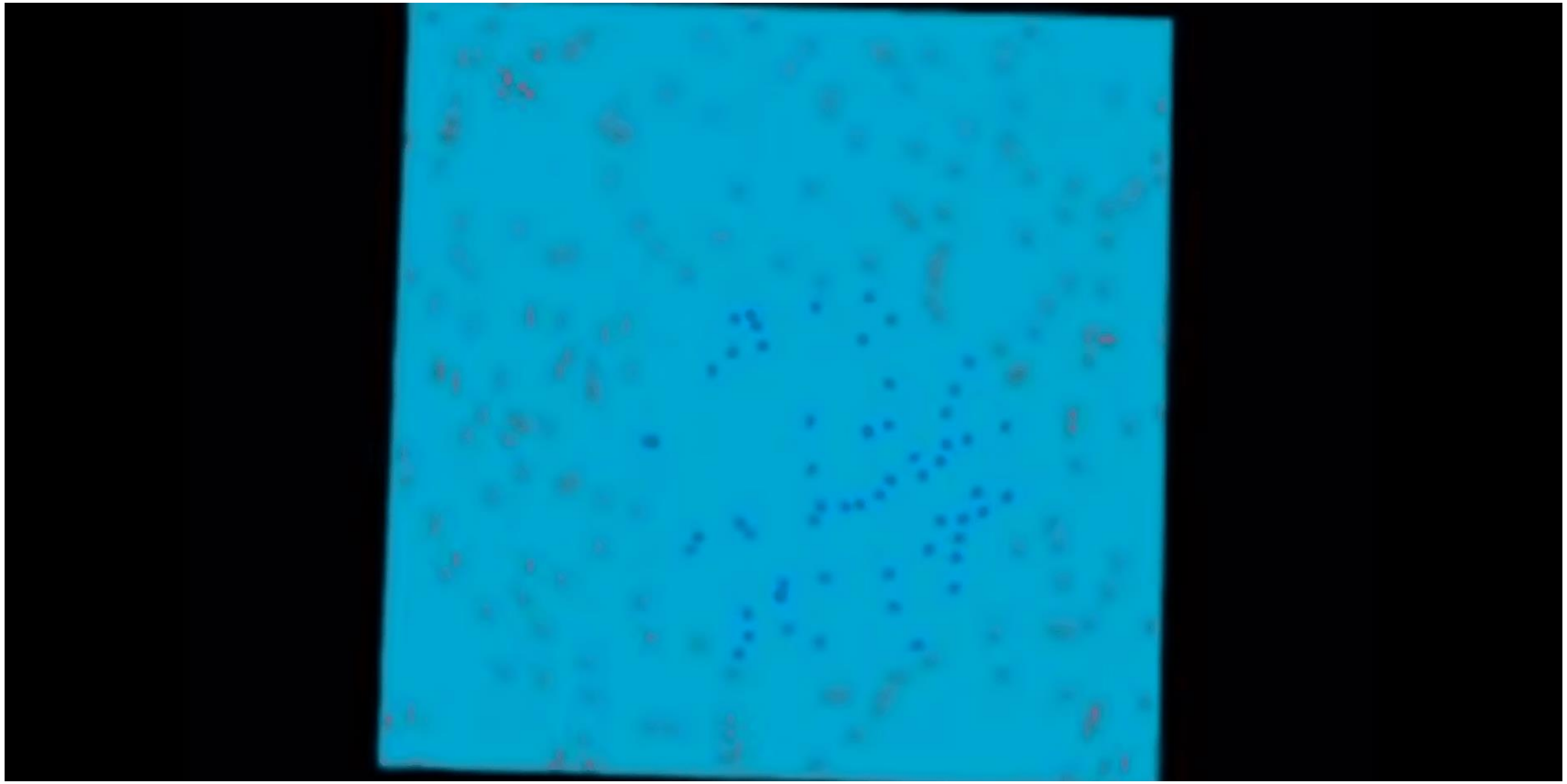
MANIPULAÇÃO DOS DADOS DE ENTRADA

MODELO LINEARMENTE SEPARÁVEIS SÃO LIMITADOS

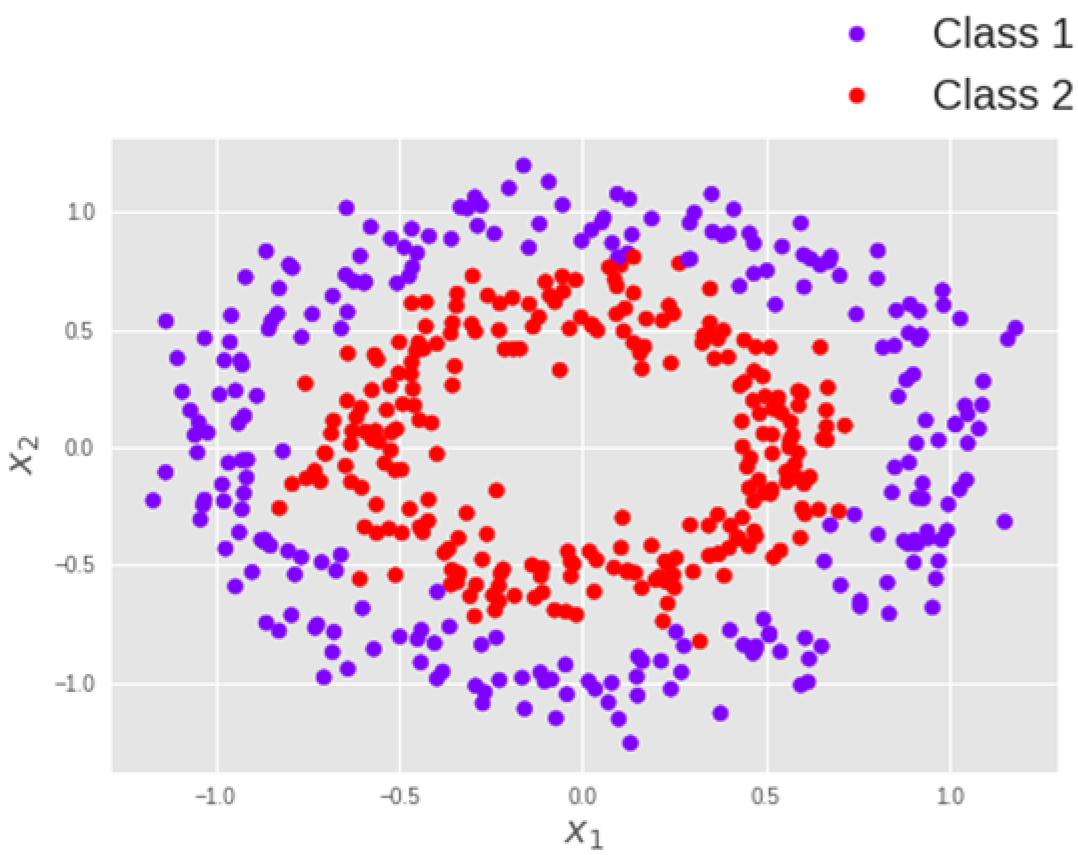




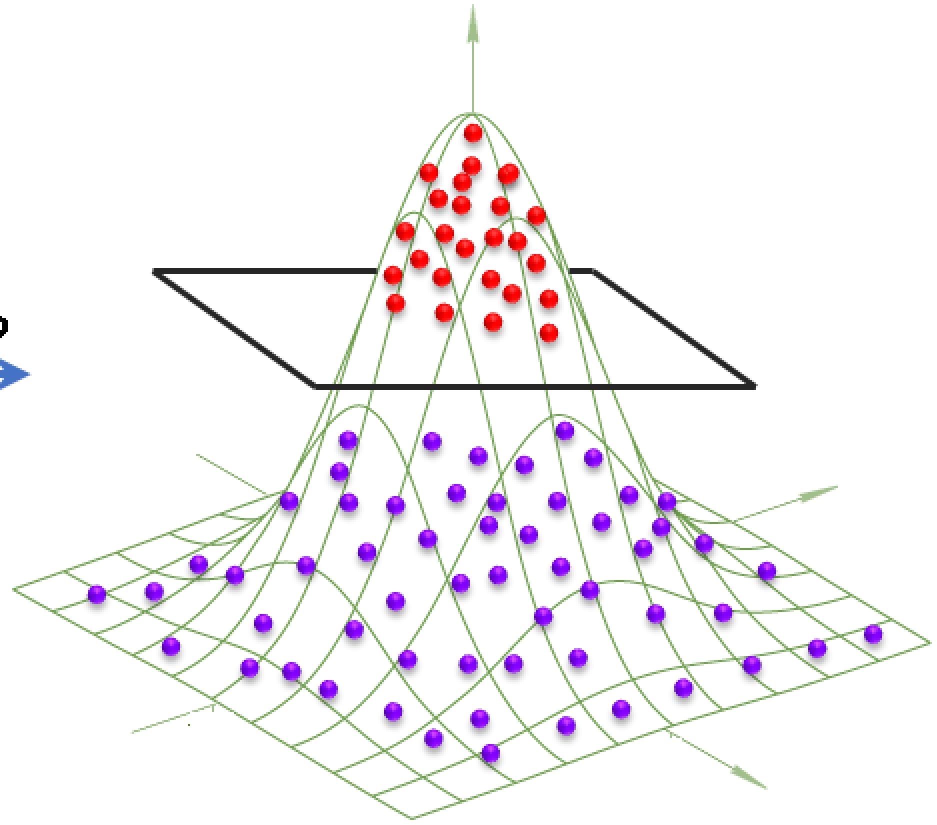
VÍDEO



https://youtu.be/3IiCbRZPrZA



$Kernel\ \phi$



LEMBRE-SE DA AULA DE REGRESSÃO...

Regressão Linear Simples

$$\hat{y} = \omega_0 + \omega_1 x$$

LINEARIDADE ESTÁ NOS PESOS ω
x SÃO CONSTANTES, DADOS DE ENTRADA.

Regressão Linear Múltipla

$$\hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \cdots + \omega_n x_n$$

Regressão Linear Polinomial

$$\hat{y} = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n$$

CONCLUSÃO: ISSO ABRE UMA FANTÁSTICA POSSIBILIDADE!

Agora, podemos pegar os dados de entrada, que são apenas constantes! Podemos fazer transformações incredíveis com esses dados...
E isso apenas se torna dados mais elaborados, mas constantes!

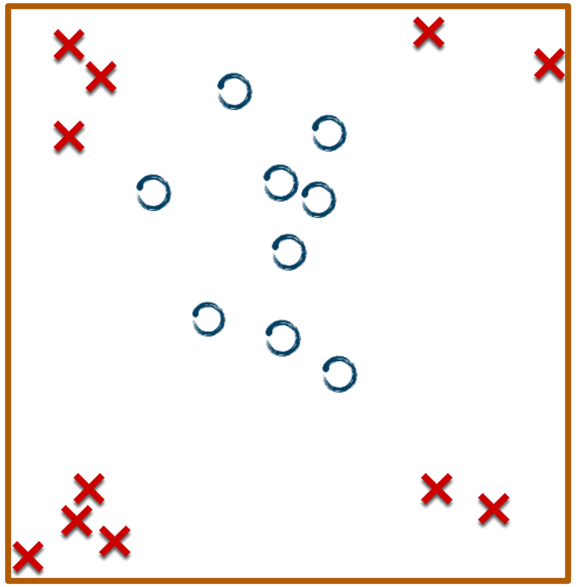
Quando vou *aprender* usando meus dados não linearmente transformados, ainda estou em um modelo linear, porque os **pesos** que serão dados às características não lineares, têm uma dependência linear!

TRANSFORMANDO DADOS

$z_n = \phi(x_n)$

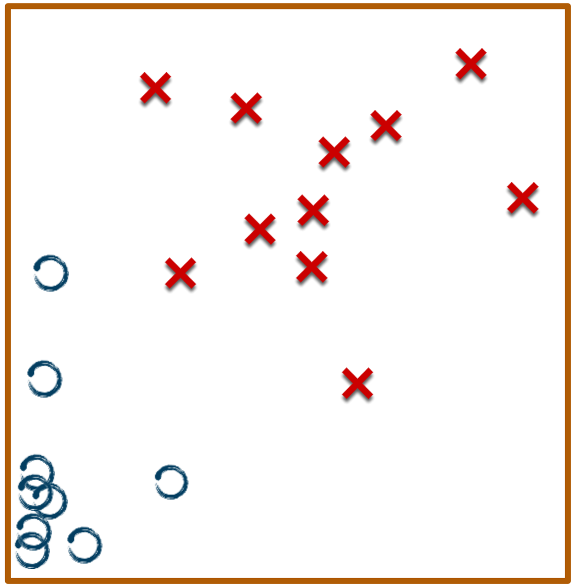
x_1, x_2

$z_1 = x_1^2, z_2 = x_2^2$



$\mathcal{X}$

$\mathcal{Z}$



VOLTANDO AO NOSSO PROBLEMA DE MINIMIZAÇÃO...

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \; {\mathbf{z}^{(i)}}^T \mathbf{z}^{(j)}$$

Lembre-se que esse espaço $\mathcal{Z}$ pode ter a dimensão que eu quiser...

Portanto, a pergunta é:
Se aumentarmos muito a nossa dimensão no espaço que criamos, qual o preço que pagamos em termos de treinamento???

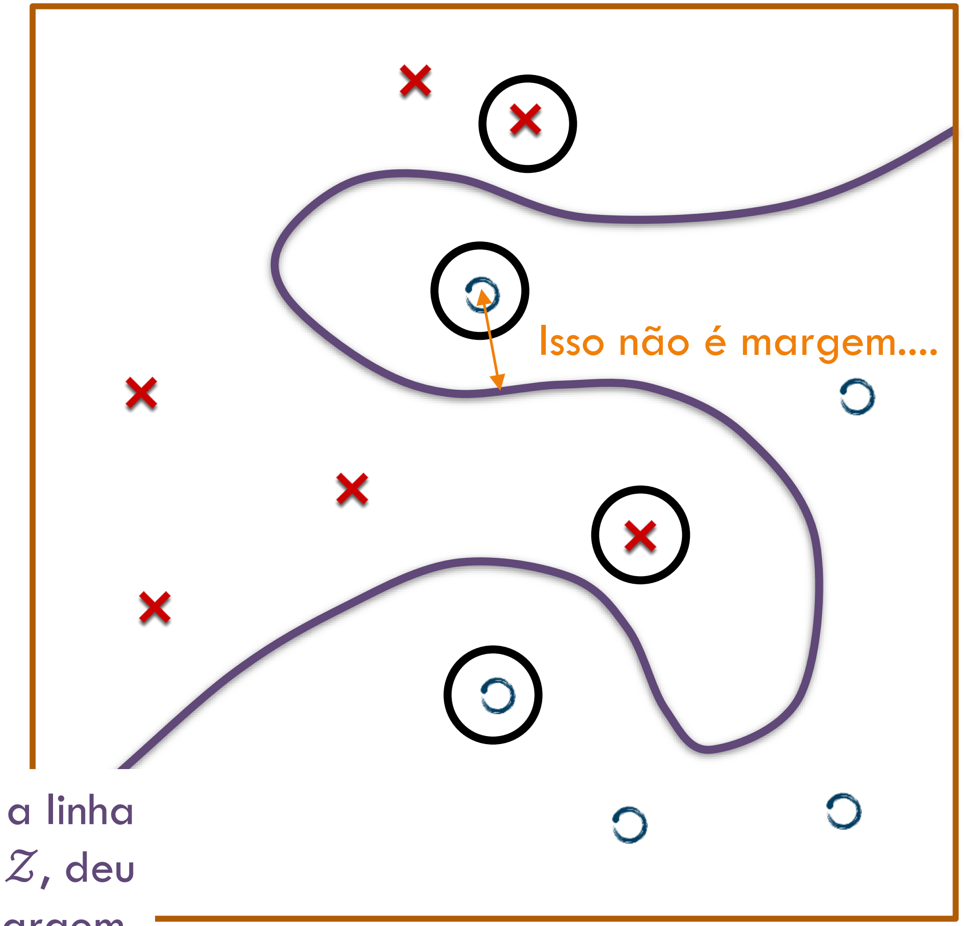
Para encontramos os α_i, temos que calcular uma quantidade que depende apenas do produto interno entre os pontos $\boldsymbol{x}$ no conjunto de treinamento. Portanto a dimensão do problema, dimensão do vetor $\boldsymbol{\alpha}$, **não muda.** Veja que o somatório tem dimensão m. **Portanto, pode-se ir a um espaço enorme, sem pagar o preço por isto em termos da otimização que será feita.**

VETORES DE SUPORTE NO ESPAÇO $\mathcal{Z}$

Vetores de suporte, margens,... existem no espaço $\mathcal{Z}$

No espaço $\mathcal{X}$ há uma interpretação...

Tenho somente 4 "*vetores de suporte*", ié, 4 parâmetros expressando ω no espaço $\mathcal{Z}$.



Simplesmente a linha que, no espaço $\mathcal{Z}$, deu a melhor margem.

TRANSFORMANDO **O QUE** EM **O QUE**?

$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$	$\xrightarrow{\phi}$	$\boldsymbol{z} = (z_1, z_2, \dots, z_{\tilde{n}})$
$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m$	$\xrightarrow{\phi}$	$\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_m$
$y_1, y_2, \dots, y_m$	$\xrightarrow{\phi}$	$y_1, y_2, \dots, y_m$
Não existem pesos em $\mathcal{X}$		$\widetilde{\boldsymbol{\omega}} = (\omega_1, \omega_2, \dots, \omega_{\tilde{n}}) = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{z}^{(i)}$, $\tilde{b} \rightarrow y^{(i)}\left(\widetilde{\boldsymbol{\omega}}^T \boldsymbol{z}^{(i)} + \widetilde{\boldsymbol{b}}\right) = \mathbf{1}$
$g(\boldsymbol{x})$	$=$	$\operatorname{sign}\left(\widetilde{\boldsymbol{\omega}}^T \phi(\boldsymbol{x}) + \widetilde{\boldsymbol{b}}\right)$

PECE Programa de Educação Continuada
Escola Politécnica da USP

ERROS

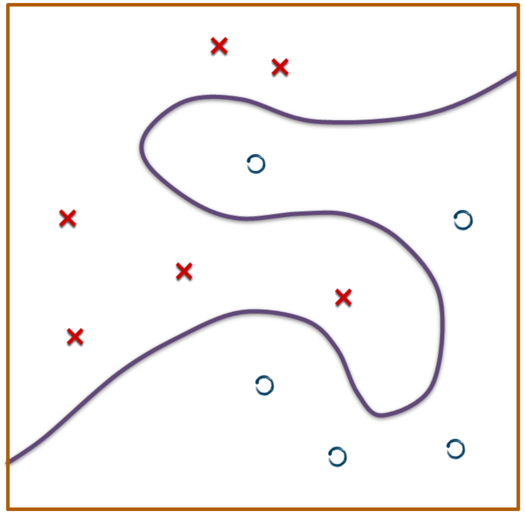
FATO...

O limite da margem máxima também possui a curiosa propriedade de que a solução depende apenas de um subconjunto dos exemplos, aqueles que aparecem exatamente na margem , i.é, os vetores de suporte. O restante dos exemplos pode estar em qualquer lugar fora da margem sem afetar a solução.

Obteríamos o mesmo classificador se tivéssemos recebido apenas os vetores de suporte como exemplos de treinamento. Isso é uma coisa boa? Como medir o erro que podemos cometer na generalização desse problema?

LIMITE SUPERIOR DE ERRO NA VALIDAÇÃO SVM

$$E_{out} \leq \frac{\#SV}{m-1}$$



$E = \frac{10}{1000}$ excelente!

$E = \frac{500}{1000}$ o-ho! Problemas...

PECE Programa de Educação Continuada
Escola Politécnica da USP

TRUQUE DO KERNEL

A IDEIA É GENIAL...

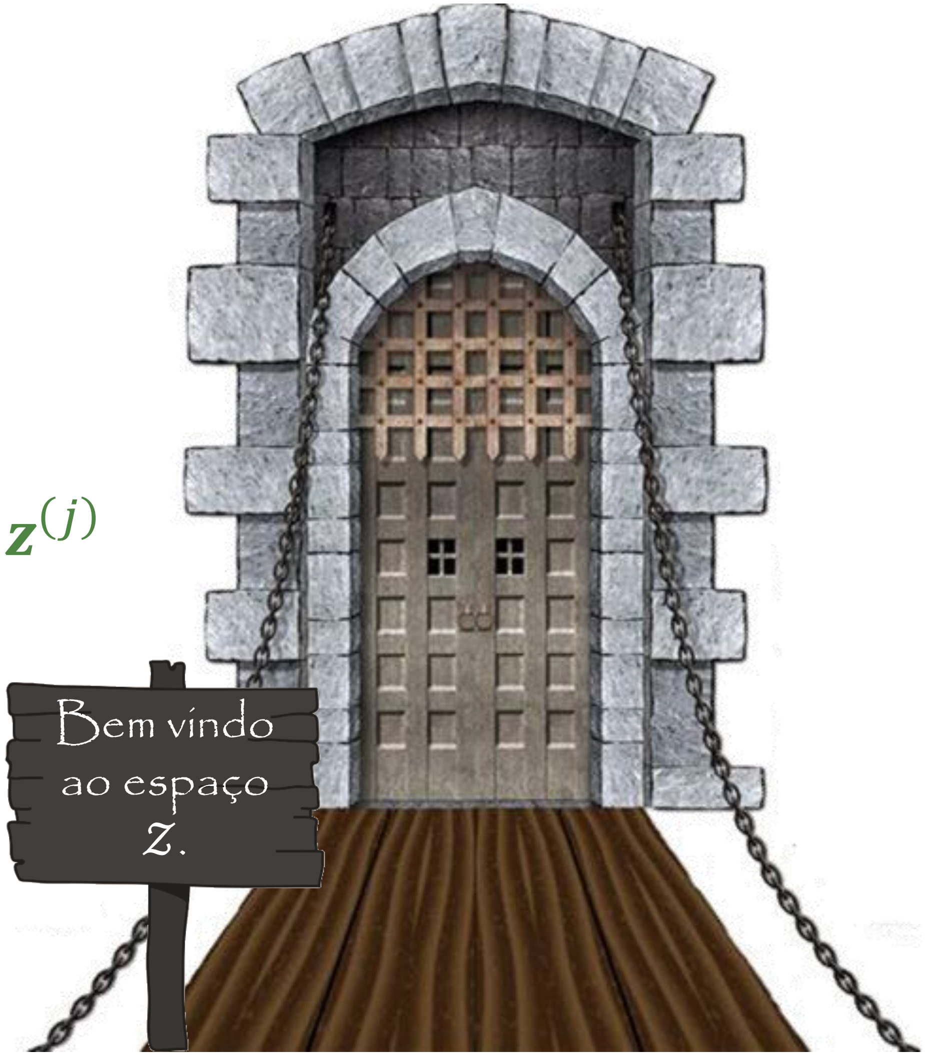
Pergunta: o que eu preciso do espaço $\mathcal{Z}$ para meu aprendizado?

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \, \mathbf{z}^{(i)^T} \mathbf{z}^{(j)}$$

A dimensionalidade de $\mathbf{z}$ não aparece explicitamente na expressão,

mas temos que realizar o produto interno $\mathbf{z}^{(i)^T}\mathbf{z}^{(j)}$ no espaço $\mathcal{Z}$.

$$\alpha_i \geq 0, i = 1, \dots, m \qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$



MAS, AINDA PRECISO...

$$g(\boldsymbol{x}) = \operatorname{sign}\left(\widetilde{\boldsymbol{\omega}}^T \boldsymbol{z} + \widetilde{\boldsymbol{b}}\right)$$

$$\widetilde{\boldsymbol{\omega}} = (\omega_1, \omega_2, \dots, \omega_{\tilde{n}}) = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{z}^{(i)},$$

$$g(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{z}^{(i)^T} \boldsymbol{z} + \widetilde{\boldsymbol{b}}\right)$$

$\tilde{b} \rightarrow y^{(j)}\left(\widetilde{\boldsymbol{\omega}}^T \boldsymbol{z}^{(j)} + \widetilde{\boldsymbol{b}}\right) = 1, j$ em SV

$$y^{(j)}\left(\sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{z}^{(i)^T} \boldsymbol{z}^{(j)} + \widetilde{\boldsymbol{b}}\right) = \mathbf{1}$$

Se sou capaz de computar o produto interno no espaço $\mathcal{Z}$, **sem conhecer $\mathcal{Z}$**, eu ainda posso treinar meu modelo!

EM RESUMO,

Dados os pontos $\boldsymbol{x}$ e $\boldsymbol{x}'$ ∈ $\mathcal{X}$, eu preciso $\boldsymbol{z}^T\boldsymbol{z}'$

Supõe-se que $\boldsymbol{z}^T\boldsymbol{z}' = K(\boldsymbol{x}, \boldsymbol{x}')$ (Kernel)

Exemplo:

$\boldsymbol{x} = (x_1, x_2) \rightarrow \phi$ segunda ordem

$\boldsymbol{z} = \phi(\boldsymbol{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$

$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{z}^T\boldsymbol{z}' = 1 + x_1 x'_1 + x_2 x'_2 + x_1^2 x'^2_1 + x_2^2 x'^2_2 + x_1 x_2 x'_1 x'_2$

QUAL É O TRUQUE?

Podemos computar $K(\boldsymbol{x}, \boldsymbol{x}')$ SEM transformar $\boldsymbol{x}$ e $\boldsymbol{x}'$?

Exemplo:

$$K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^T \boldsymbol{x}')^2$$

Isto não é um produto interno em $\mathcal{X}$, é uma função... Nem é claramente um produto interno em outro espaço transformado...

$$= (1 + x_1 x'_1 + x_2 x'_2)^2$$

$$= 1 + x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2 x_1 x'_1 + 2 x_2 x'_2 + 2 x_1 x_2 x'_1 x'_2$$

Este **É** um produto interno.

$$\left(1, x_1^2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, \sqrt{2} x_1 x_2\right)$$

$$\left(1, x'^2_1, x'^2_2, \sqrt{2} x'_1, \sqrt{2} x'_2, \sqrt{2} x'_1 x'_2\right)$$

KERNEL POLINOMIAL

$\mathcal{X} \in \mathbb{R}^n$ e $\phi: \mathcal{X} \to \mathcal{Z}$ é um polinômio de ordem q

O *kernel equivalente* é,

$$K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^T \boldsymbol{x}')^q$$
$$= (1 + x_1 x'_1 + x_2 x'_2 + \cdots + x_n x'_n)^q$$

Pode-se ajustar a escala, $K(\boldsymbol{x}, \boldsymbol{x}') = (a\boldsymbol{x}^T \boldsymbol{x}' + b)^q$

PRECISAMOS, ENTÃO, QUE Z EXISTA!

Considere que $K(\boldsymbol{x}, \boldsymbol{x}')$ um produto interno em um espaço $\mathcal{Z}$, então estamos ok!

Exemplo:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$$

Função de $\boldsymbol{x}$ e $\boldsymbol{x}$', mas nem sequer aparece o produto interno explicitamente.

PRECISAMOS QUE EXISTA UMA TRANSFORMAÇÃO EM z, SEM NEM AO MENOS IMAGINAR z

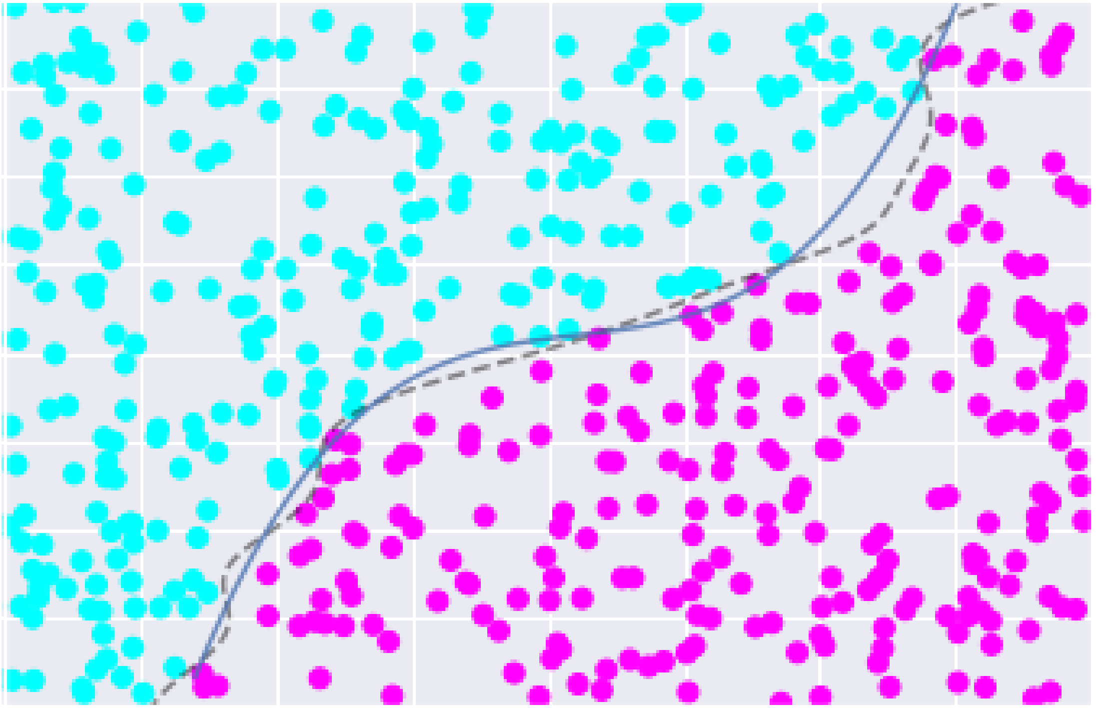
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$$

Função de Base Radial (RBF)

Pode-se provar que essa função é equivalente a pegar $\boldsymbol{x}$ e $\boldsymbol{x}$' e transformar para o espaço $\mathcal{Z}$ e extrair o produto interno dos vetores transformados $\boldsymbol{z}$ e $\boldsymbol{z}'$?

Pois, pode-se provar sim, e mais, que esse espaço $\mathcal{Z}$ tem dimensão infinita!

VAMOS PARA UM EXEMPLO...



Exagero ir até o infinito???

Cheque o número de vetores de suporte!!! Esse é o seu guia!

```
Número de vetores de suporte:  16
Graus de liberdade:  500
Limite erro:  0.032
```

FORMALIDADE: COMO FORMULAR O PROBLEMA?

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \underbrace{\begin{bmatrix} y^{(1)}y^{(1)}K\left(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(1)}\right) & y^{(1)}y^{(2)}K\left(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}\right) & \cdots & y^{(1)}y^{(m)}K\left(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(m)}\right) \\ y^{(2)}y^{(1)}K\left(\boldsymbol{x}^{(2)},\boldsymbol{x}^{(1)}\right) & y^{(2)}y^{(2)}K\left(\boldsymbol{x}^{(2)},\boldsymbol{x}^{(2)}\right) & \cdots & y^{(2)}y^{(m)}K\left(\boldsymbol{x}^{(2)},\boldsymbol{x}^{(m)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ y^{(m)}y^{(1)}K\left(\boldsymbol{x}^{(m)},\boldsymbol{x}^{(1)}\right) & y^{(m)}y^{(2)}K\left(\boldsymbol{x}^{(m)},\boldsymbol{x}^{(2)}\right) & \cdots & y^{(m)}y^{(m)}K\left(\boldsymbol{x}^{(m)},\boldsymbol{x}^{(m)}\right) \end{bmatrix}}_{\text{Coeficientes quadráticos}} \boldsymbol{\alpha} + \underbrace{(-1 \quad \cdots \quad -1)}_{\text{Linear}}\boldsymbol{\alpha}$$

sujeito a

$$\underbrace{\boldsymbol{y}^T\boldsymbol{\alpha} = 0}_{\text{Restrição Linear}}$$

$$\underbrace{\boldsymbol{0} \leq \boldsymbol{\alpha} \leq \infty}_{\text{Limites inferior e superior}}, i = 1, \dots, m$$

O que o problema de otimização devolve???? $\boldsymbol{\alpha}$

O que a biblioteca em python devolve???? $\tilde{b}, \tilde{\omega}$

A HIPÓTESE FINAL

$g(\boldsymbol{x}) = \text{sign}\left(\widetilde{\boldsymbol{\omega}}^T \boldsymbol{z} + \widetilde{\boldsymbol{b}}\right)$ em termos do kernel K

$$g(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{SV} \alpha_i y^{(i)} K\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}\right) + \widetilde{\boldsymbol{b}}\right)$$

Este é o seu modelo!!!

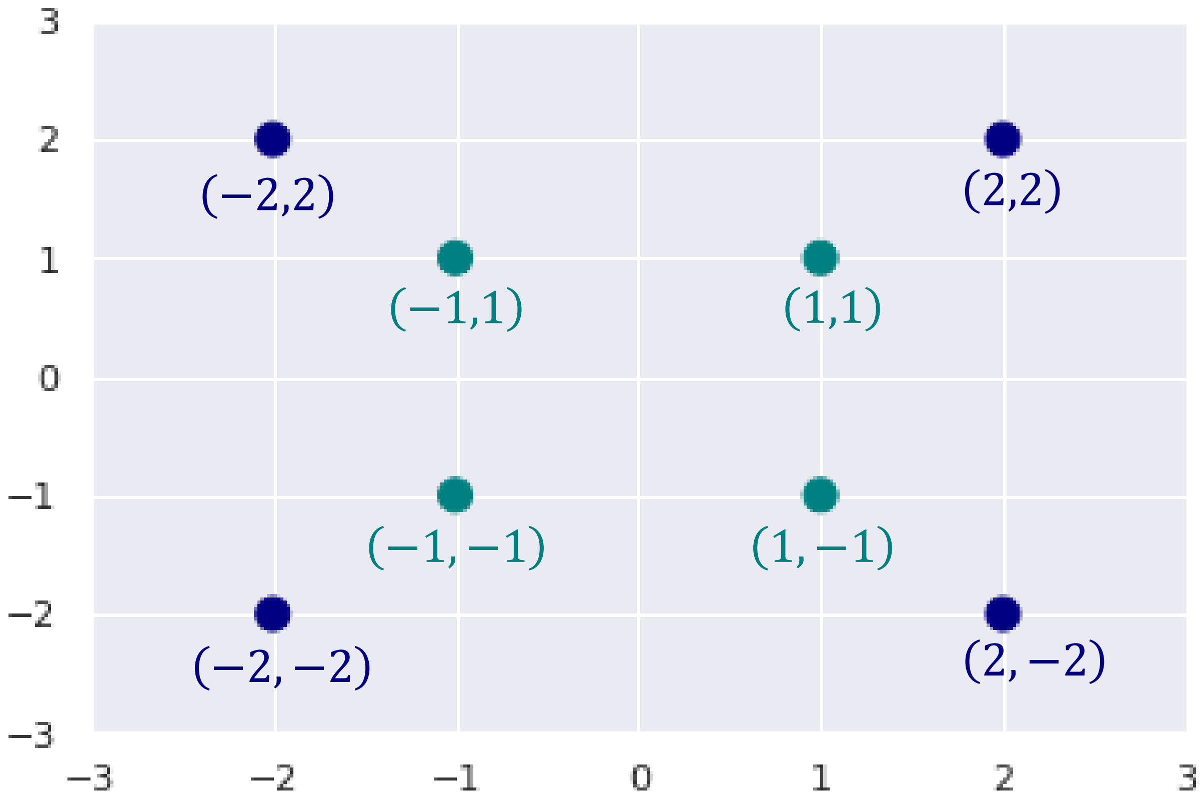
$\tilde{b} \rightarrow y^{(j)}\left(\widetilde{\boldsymbol{\omega}}^T \boldsymbol{z}^{(j)} + \widetilde{\boldsymbol{b}}\right) = 1, j$ em SV

$$\widetilde{\boldsymbol{b}} = y^{(j)} - \sum_{i=1}^{SV} \alpha_i y^{(i)} K\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)$$

EXEMPLO NOTEBOOK

$$x_+ = \left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \end{pmatrix} \right\}$$

$$x_- = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$$

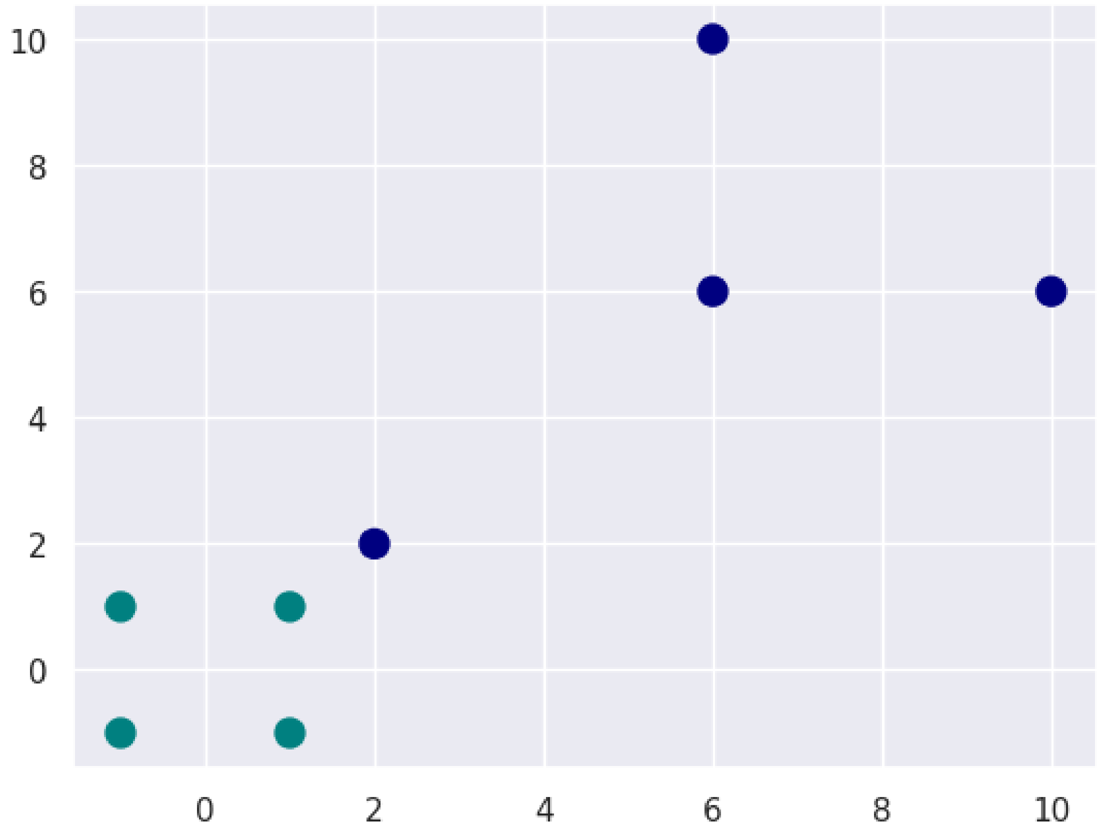


$$z = \phi(x_1, x_2) = \begin{cases} (4 - x_2 + |x_1 - x_2|, 4 - x_1 + |x_1 - x_2|) & \text{se } \sqrt{x_1^2 + x_2^2} > 2 \\ (x_1, x_2) & c.c. \end{cases}$$

$$z_+ = \left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 10 \\ 6 \end{pmatrix}, \begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 6 \\ 10 \end{pmatrix} \right\}$$

$$z_- = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$$

$$sv_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad sv_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

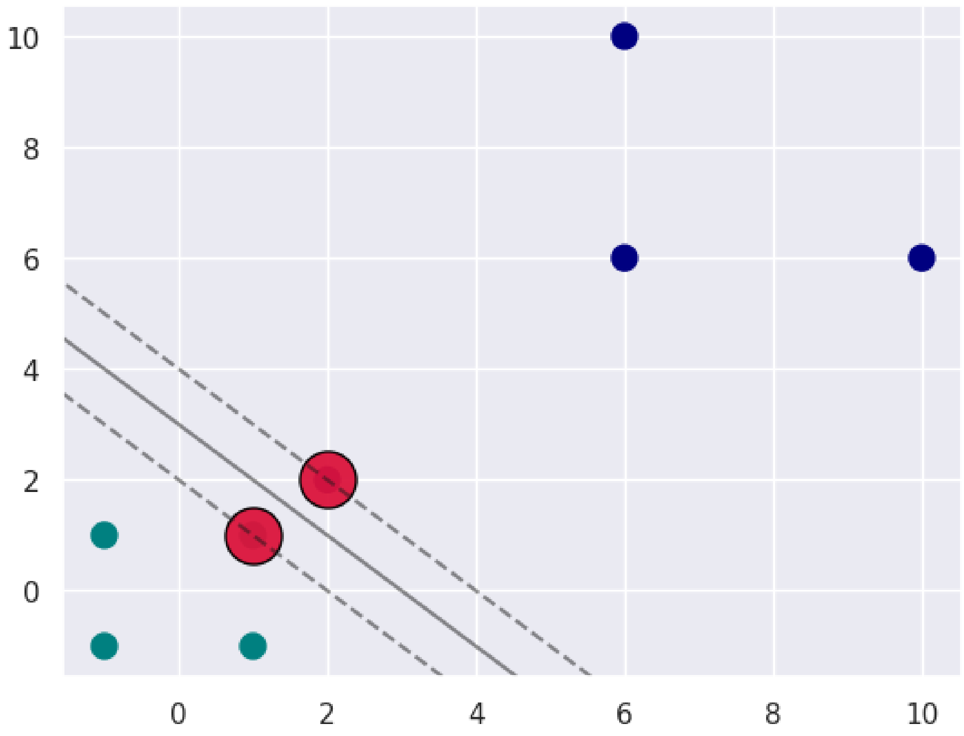


$$
\begin{aligned}
-\alpha_1 z^{(1)} \cdot z^{(1)} + \alpha_2 z^{(2)} \cdot z^{(1)} + \bar{b} &= -1 \\
-\alpha_1 z^{(1)} \cdot z^{(2)} + \alpha_2 z^{(2)} \cdot z^{(2)} + \bar{b} &= +1 \\
-\alpha_1 + \alpha_2 &= 0
\end{aligned}
$$

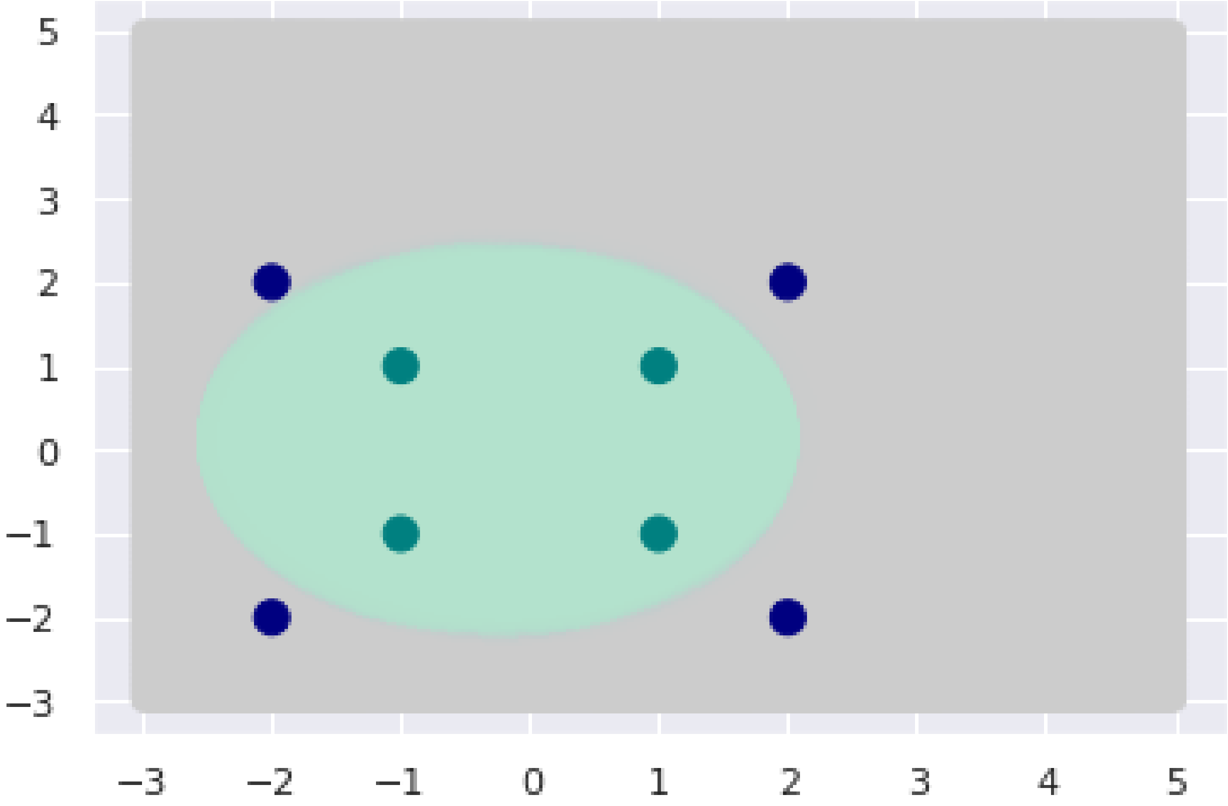
$$
\begin{aligned}
-2\alpha_1 + 4\alpha_2 + \bar{b} &= -1 \\
-4\alpha_1 + 8\alpha_2 + \bar{b} &= +1 \\
-\alpha_1 + \alpha_2 + 0\bar{b} &= 0
\end{aligned}
$$

$$
\bar{\omega} = \sum_{i=1}^{n_{sv}} \alpha_i y^{(i)} z^{(i)}
$$

```
alpha 1 = 1.00, alpha 2 = 1.00
b = -3.00
w 1 = 1.00, w 2 = 1.00
```



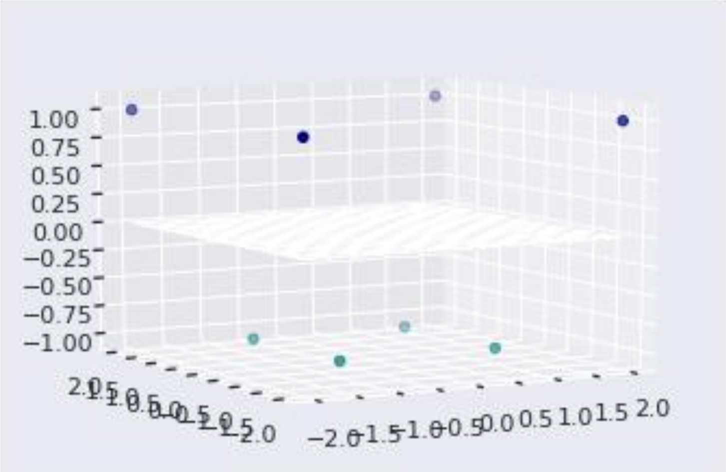
SE EU RETORNO A x



KERNEL

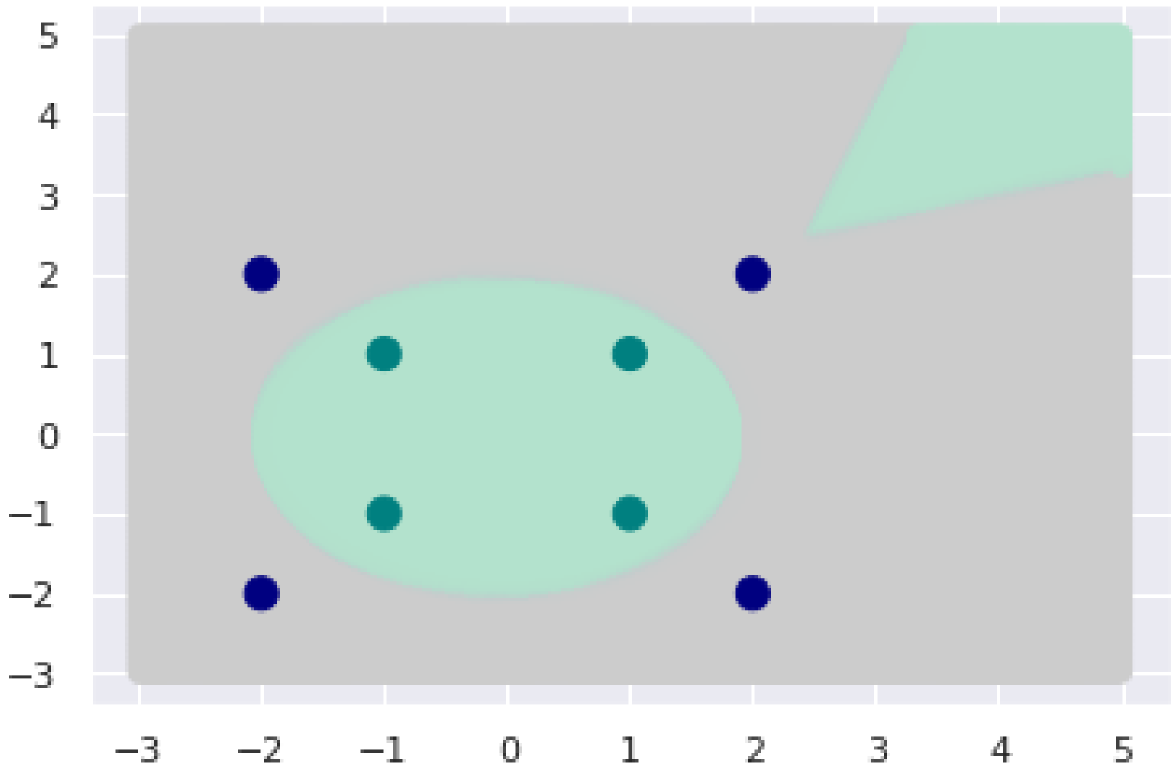
$$z = \phi_2(x_1, x_2) = \begin{pmatrix} x_1 \\ x_2 \\ \frac{(x_1^2+x_2^2)-5}{3} \end{pmatrix}$$

$$z_+ = \left\{ \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} \right\}$$



$$z_- = \left\{ \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} \right\}$$

SE EU RETORNO A x



PECE Programa de Educação Continuada
Escola Politécnica da USP

KERNEL VÁLIDO

Existem 3 métodos para obtenção de um Kernel válido,

1. Construindo, como foi feito com polinômio

$$(1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2)$$
$$(1, x'^2_1, x'^2_2, \sqrt{2}x'_1, \sqrt{2}x'_2, \sqrt{2}x'_1x'_2)$$

2. Através de propriedades matemáticas (*Condição de Mercer)*

$K(\boldsymbol{x}, \boldsymbol{x}')$ é válido se, e somente se,
1. É simétrico $K(\boldsymbol{x}, \boldsymbol{x}')= K(\boldsymbol{x}', \boldsymbol{x})$ e 2. A matriz

$$\begin{bmatrix} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(m)}) \\ K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(m)}) \end{bmatrix}$$

positiva semi-definida para qualquer $\boldsymbol{x}^{(1)} \cdots \boldsymbol{x}^{(m)}$.

3. *Dakara Nani*

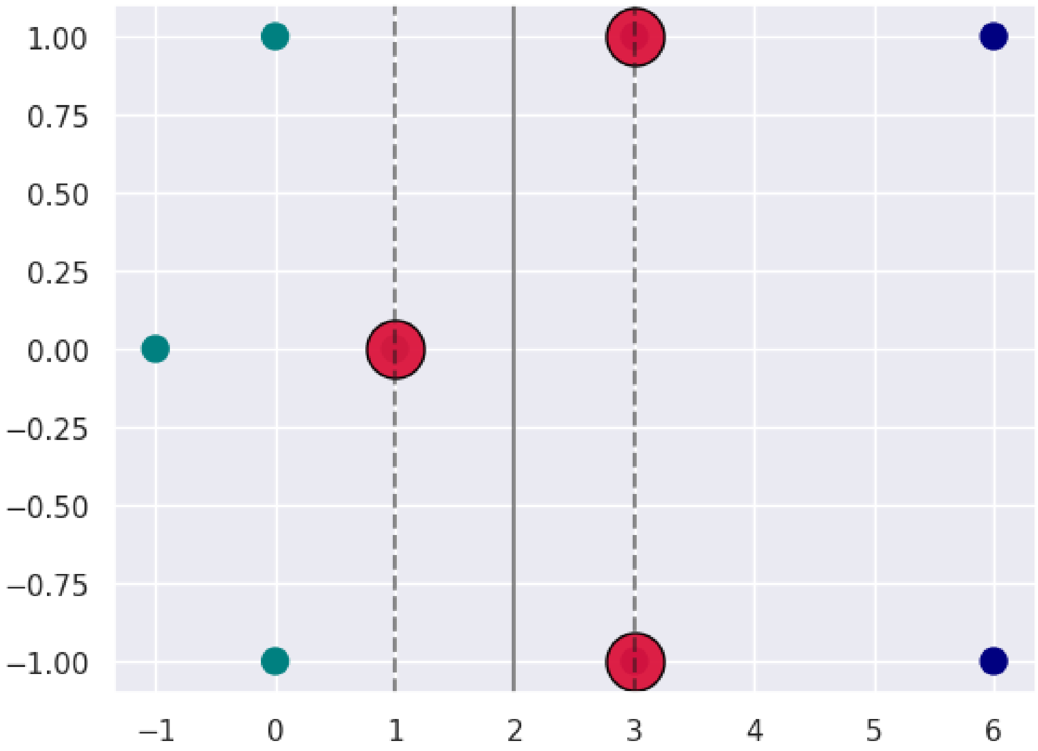
KERNELS MAIS POPULARES

- Kernel Polinomial;
- Kernel Gaussiano;
- Função de Base Radial (RBF);
- Kernel Laplace RBF;
- Kernel Sigmoide;
- Kernel Bessel;

Kernel function	Expression	Parameter
Linear kernel function	$K(x_i, x_j) = x_i \cdot x_j$	
Polynomial kernel function	$K(x_i, x_j) = \left(x_i \cdot x_j + 1\right)^d$	d
Radial basis function (RBF) kernel function	$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right)$	$\gamma > 0$
Sigmoid kernel function	$K\left(x_i, x_j\right) = \tanh\left(b\left(x_i, x_j\right) + c\right)$	b, c

PECE Programa de Educação Continuada
Escola Politécnica da USP

A SOLUÇÃO É SIMPLES....



Vamos passo a passo seguir nossa metodologia, para chegarmos à essa conclusão matematicamente...

```
Coeficientes do modelo
w1, w2 =[[ 1.00048166e+00 -1.66533454e-16]]
b = [-2.00112388]
```

PRIMEIRAMENTE, SEM RESOLVER A OTIMIZAÇÃO

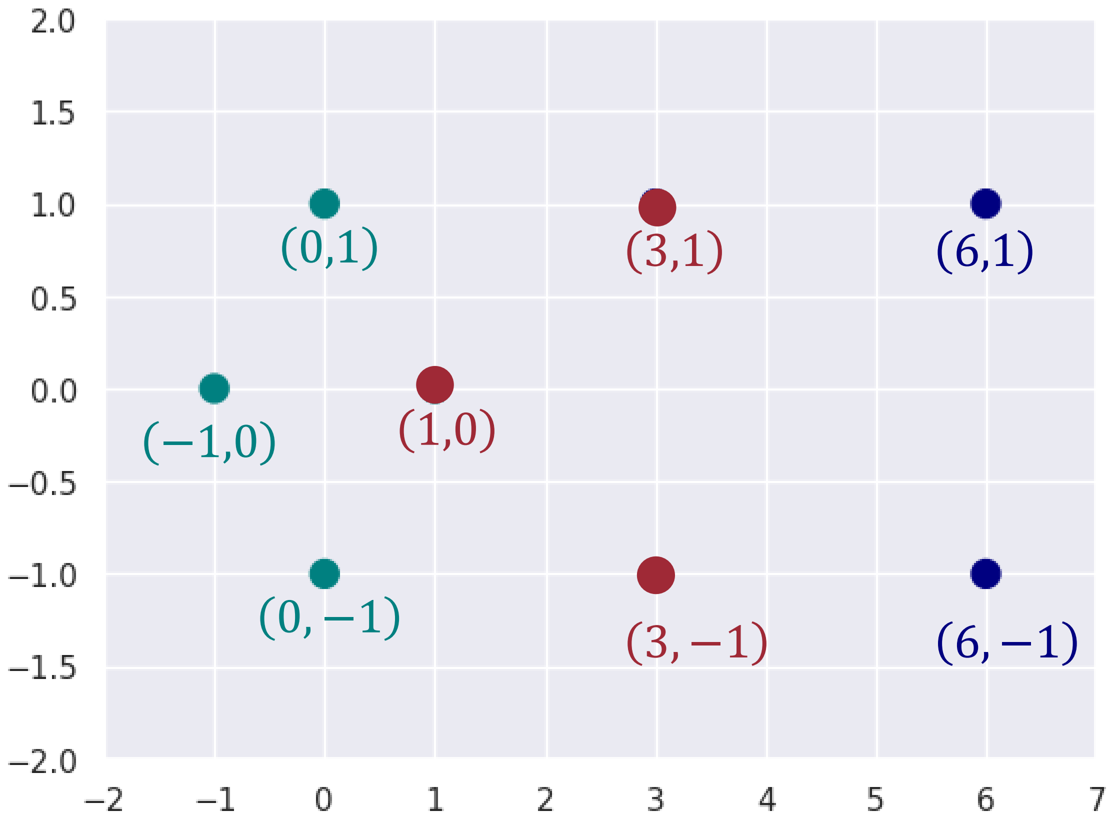
Por inspeção, verifica-se que os vetores de suporte são 3: $(1{,}0)$ para $y = -1$; $(3{,}1)$, para $y = +1$; e $(3,-1)$, para $y = +1$. Dessa forma, pode-se escrever os vetores de suporte $sv_i = 1, \cdots, 3$ como,

$$sv_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad sv_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad sv_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

Se:

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \qquad y^{(i)}\left(\boldsymbol{\omega}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

$$\omega^T x^{(j)} + b = \sum_{i=1}^{n_{sv}} \alpha_i y^{(i)} \langle x^{(i)}, x^{(j)} \rangle + b = y^{(i)}$$



$$\sum_{i=1}^{n_{sv}} \alpha_i y^{(i)} \langle x^{(i)}, x^{(j)} \rangle + b = y^{(i)}$$

$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(1)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(1)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(1)} + b = y^{(1)}$$
$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(2)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(2)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(2)} + b = y^{(2)}$$
$$\alpha_1 y^{(1)} x^{(1)} \cdot x^{(3)} + \alpha_2 y^{(2)} x^{(2)} \cdot x^{(3)} + \alpha_3 y^{(3)} x^{(3)} \cdot x^{(3)} + b = y^{(3)}$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = -\alpha_1 + \alpha_2 + \alpha_3 = 0$$

$$sv_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad sv_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad sv_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

$\langle x^{(1)}, x^{(1)} \rangle = 1$, $\langle x^{(2)}, x^{(2)} \rangle = 10$, $\langle x^{(3)}, x^{(3)} \rangle = 10$, $\langle x^{(1)}, x^{(2)} \rangle = 3$, $\langle x^{(2)}, x^{(3)} \rangle = 8$ e $\langle x^{(1)}, x^{(3)} \rangle = 3$,

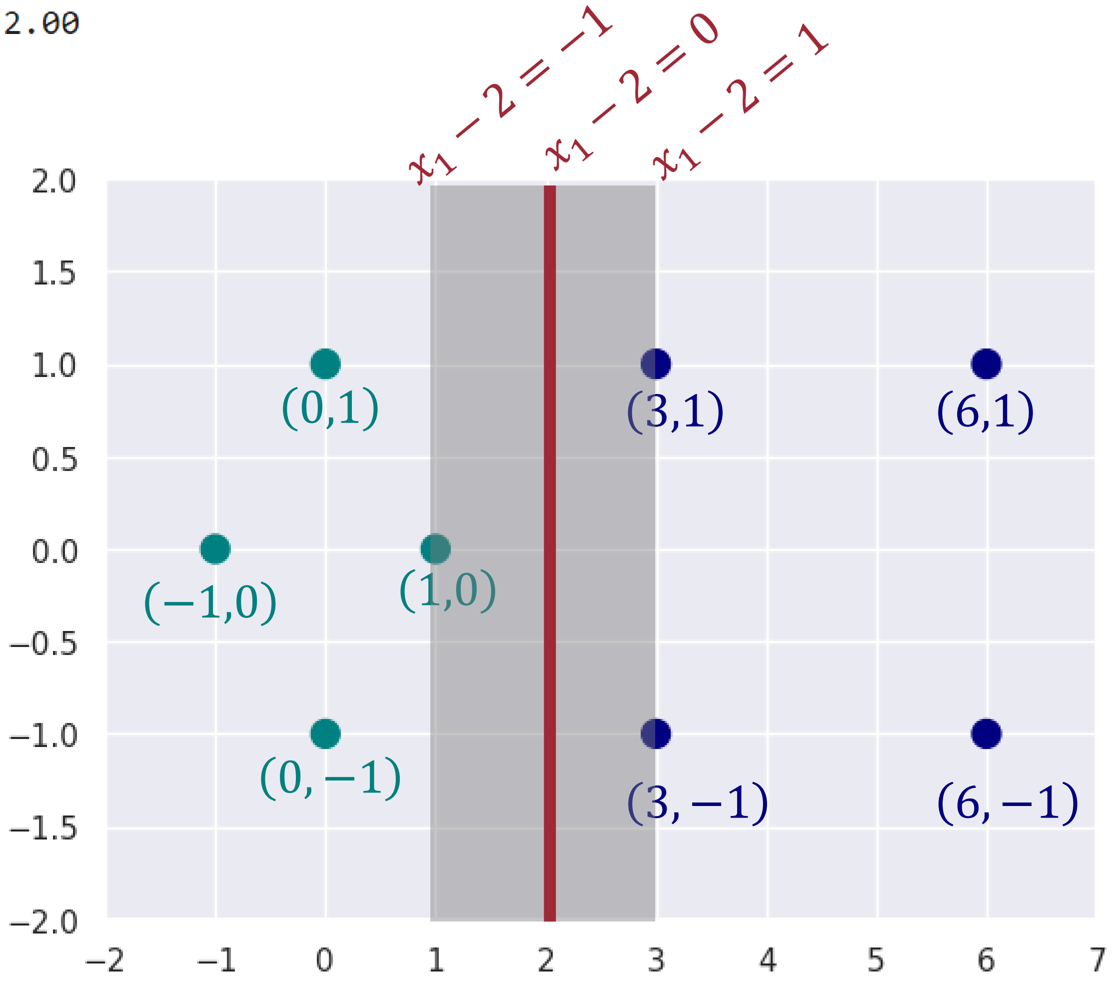
$$
\begin{aligned}
-1\alpha_1 + 3\alpha_2 + 3\alpha_3 + b &= -1 \\
-3\alpha_1 + 10\alpha_2 + 8\alpha_3 + b &= +1 \\
-3\alpha_1 + 8\alpha_2 + 10\alpha_3 + b &= +1 \\
-\alpha_1 + \alpha_2 + \alpha_3 + 0b &= 0
\end{aligned}
$$

```
[[-1.  3.  3.  1.]
 [-3. 10.  8.  1.]
 [-3.  8. 10.  1.]
 [-1.  1.  1.  0.]]
alpha 1 = 0.50, alpha 2 = 0.25, alpha 3 = 0.25
b = -2.00
```

PECE Programa de Educação Continuada
Escola Politécnica da USP

$$\boldsymbol{\omega} = \sum_{i=1}^{SV} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

```
w 1 = 1.00, w 2 = 0.00
```



PECE Programa de Educação Continuada
Escola Politécnica da USP

VOCÊ NÃO TEM DEVER DE CASA

Reveja o Notebook, existem alguns itens que você deve completar (estão com gabarito).
Se prepare para aula da próxima semana!!!