Insper

Machine Learning

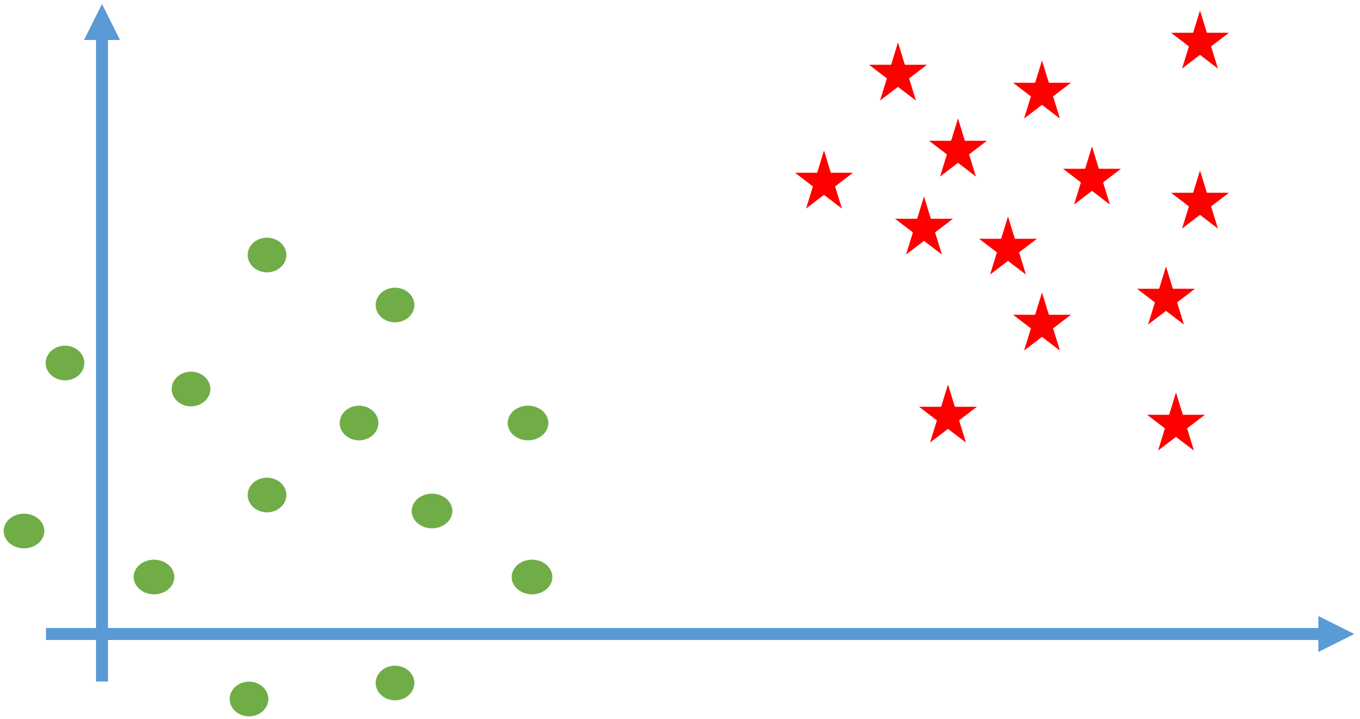
Aula 13 – Support Vector Machines

2024 – Engenharia
Fábio Ayres <fabioja@insper.edu.br>

Objetivos da aula

- Motivação para SVMs
- Hard e soft-margin
- Extensões para problemas não-lineares: kernels
- Prática

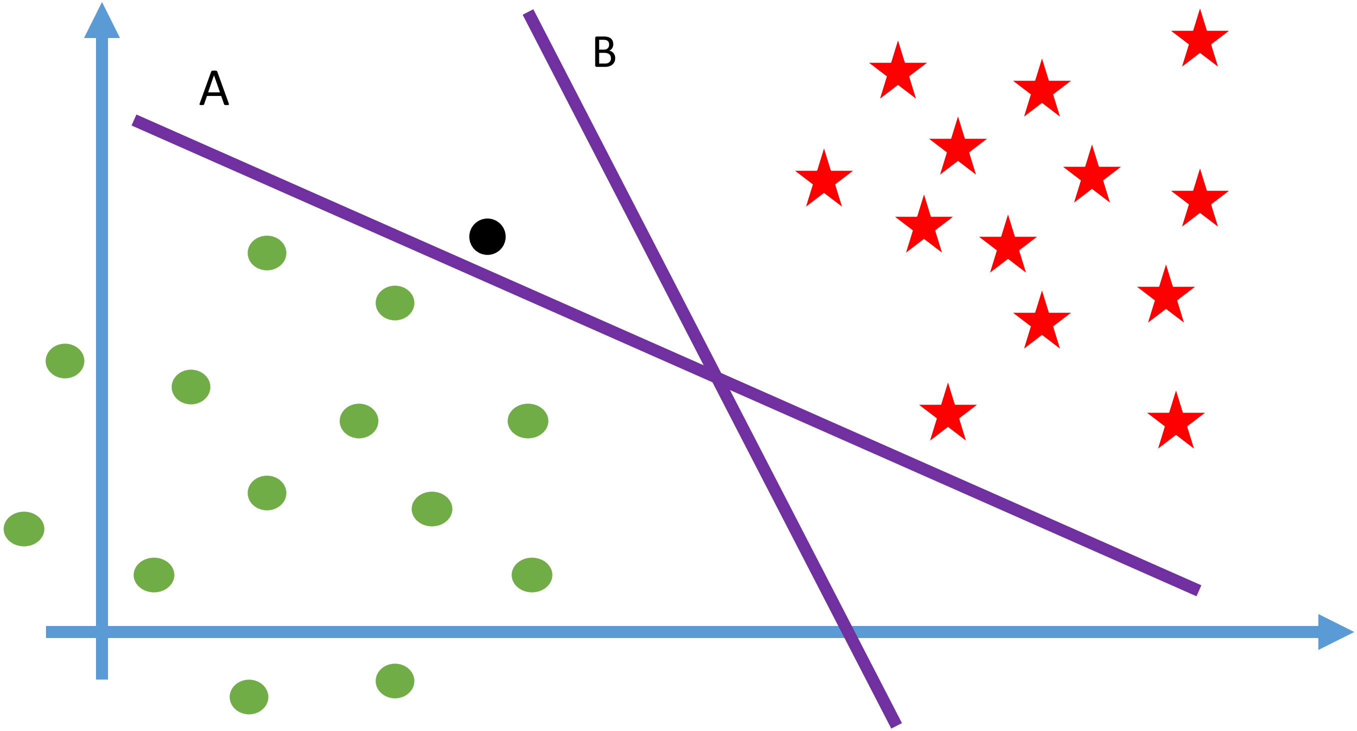
Um problema de classificação



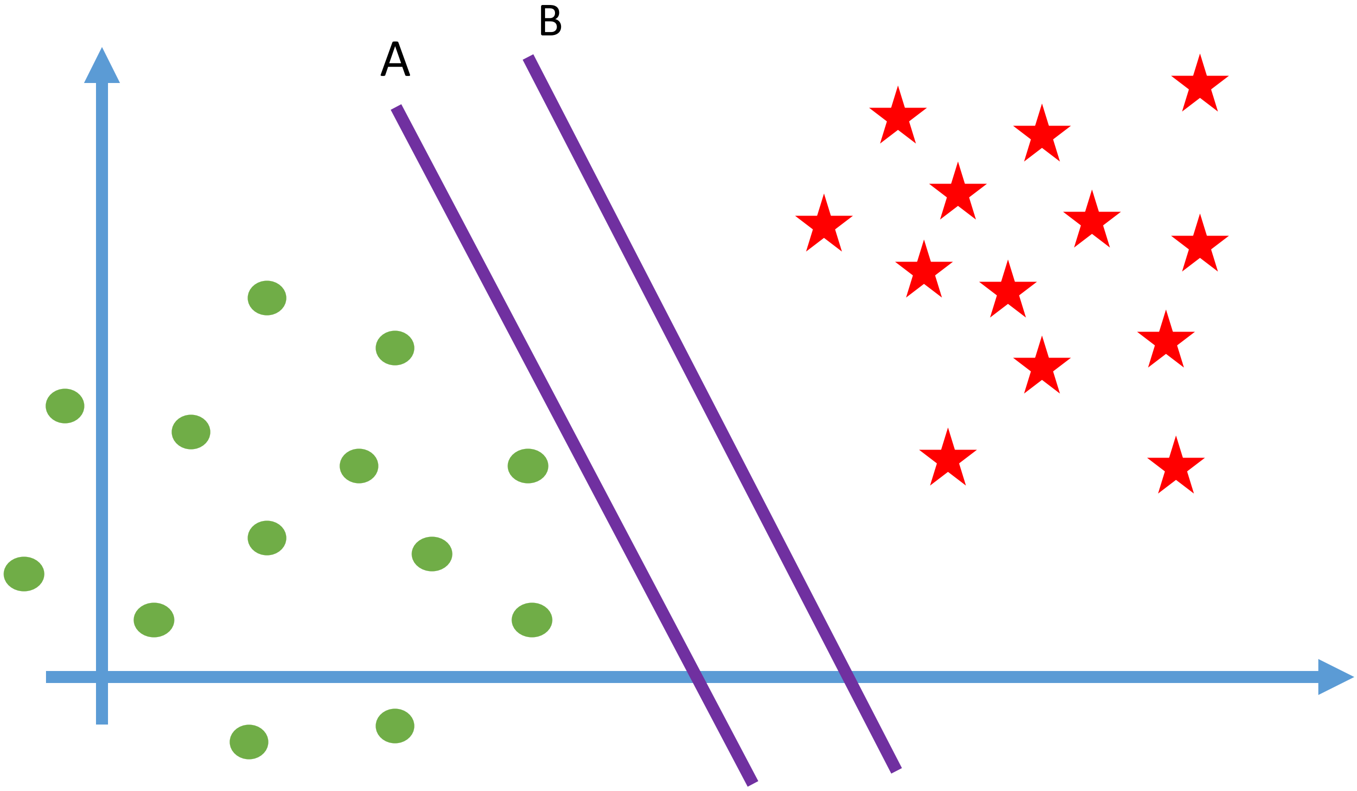
Um problema de classificação

ATENÇÃO: ESTES SÃO OS PONTOS DE TREINAMENTO!

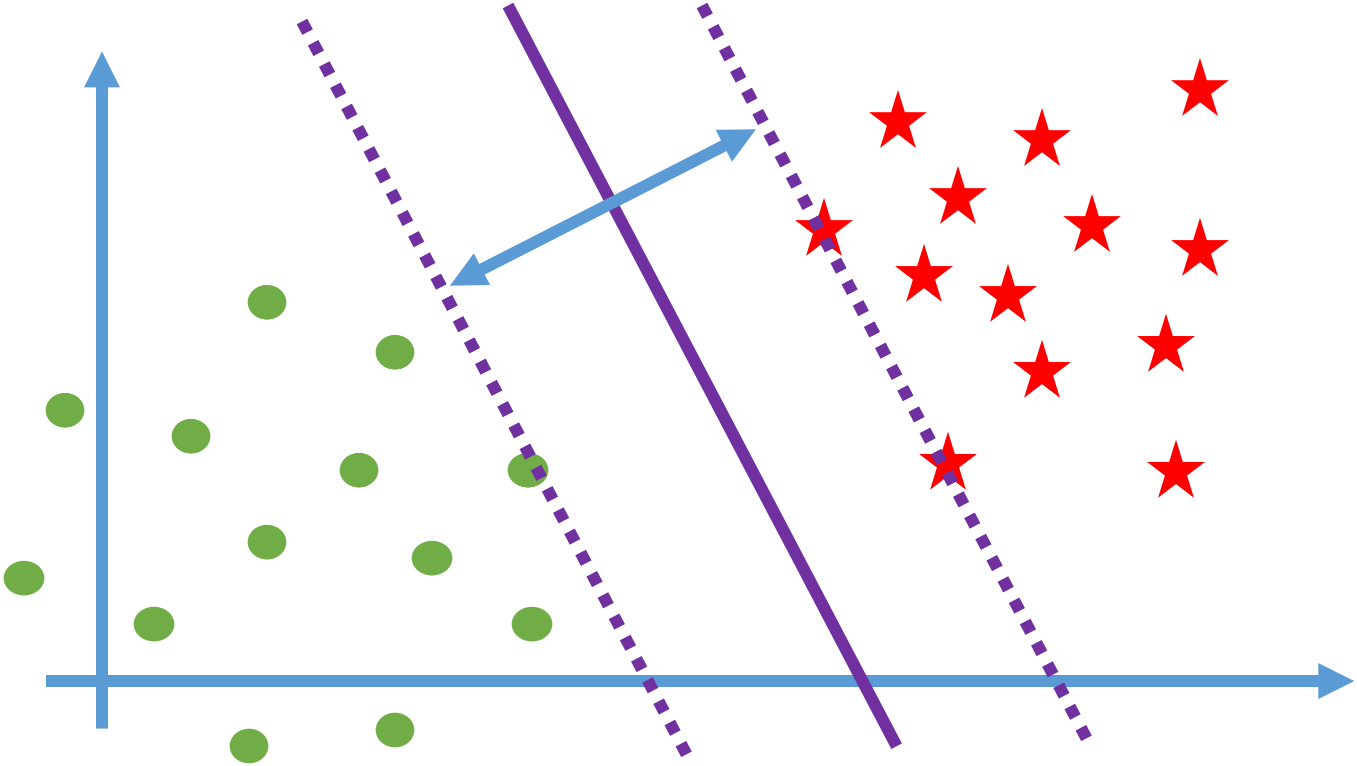
Qual a melhor reta de separação?



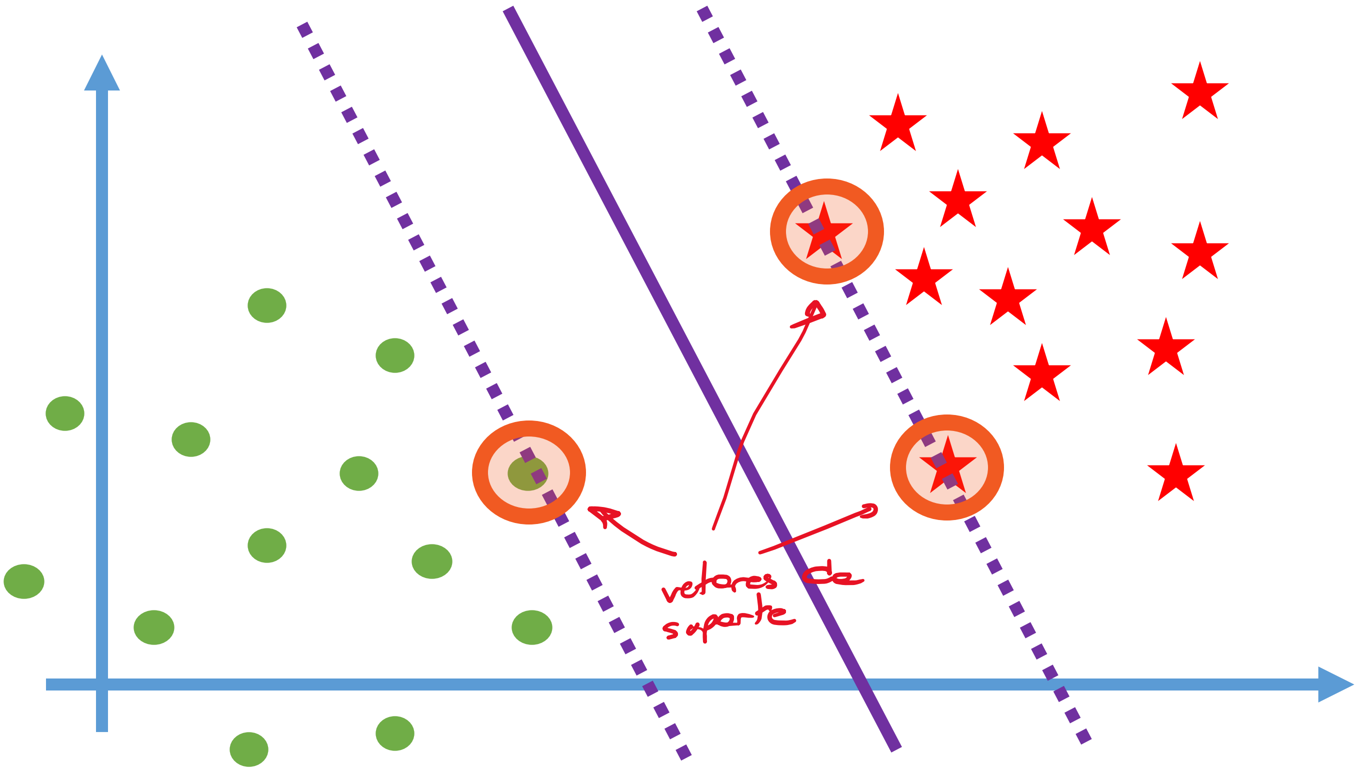
Qual a melhor reta de separação?



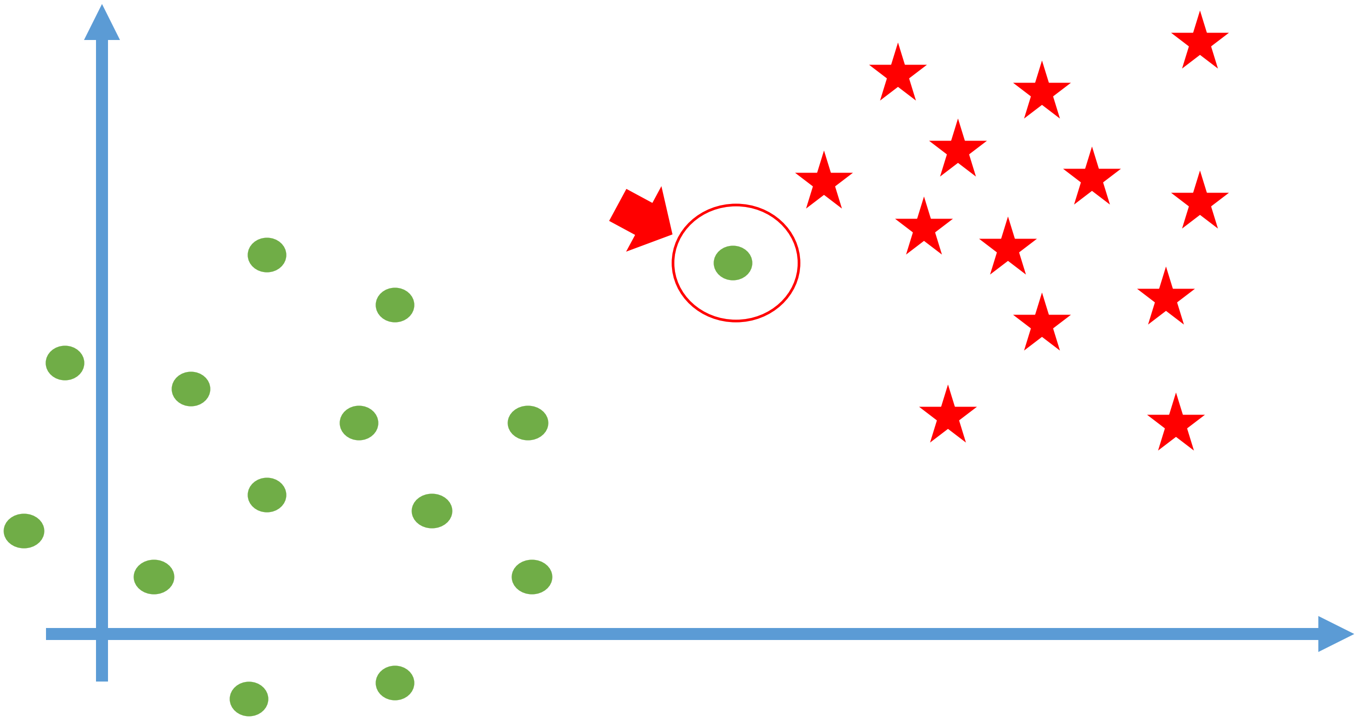
Ideia: aumentar a “avenida”



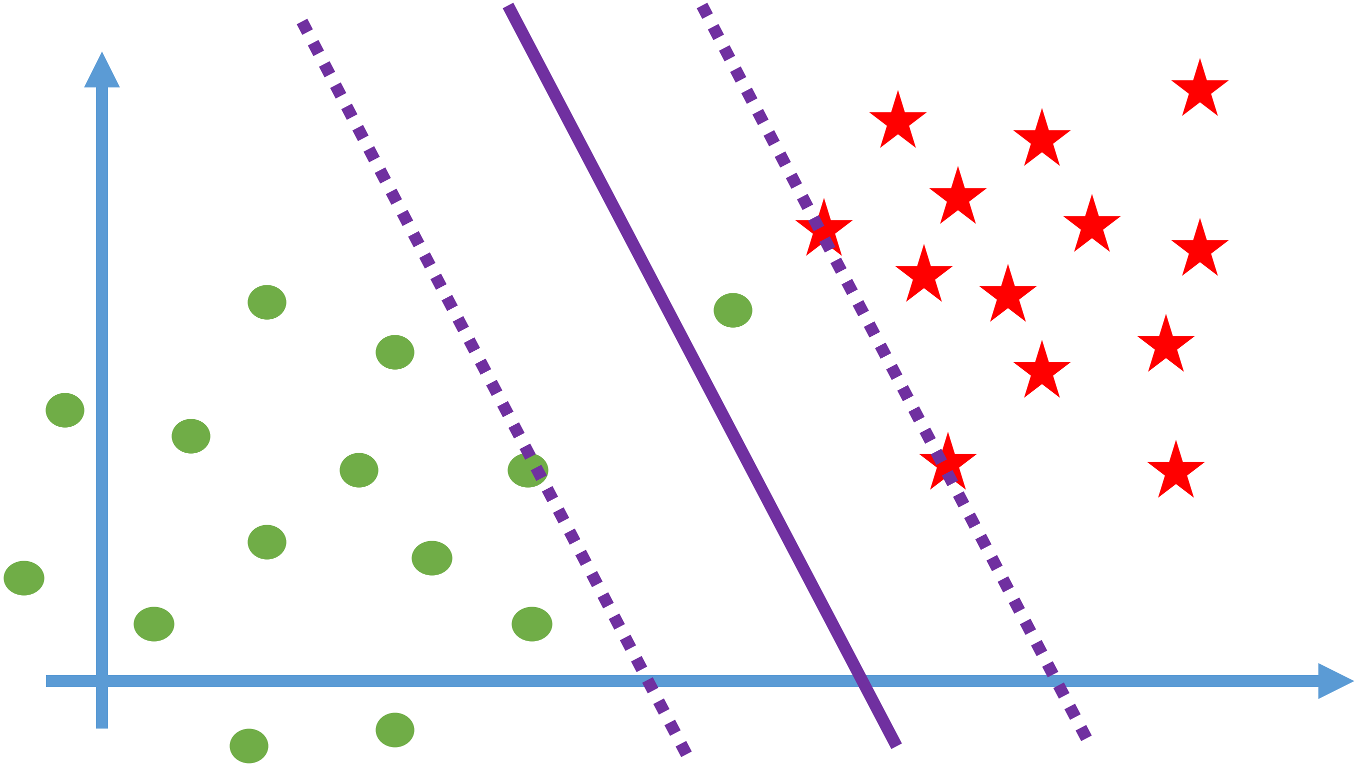
Vetores de suporte



Problemas no paraíso...

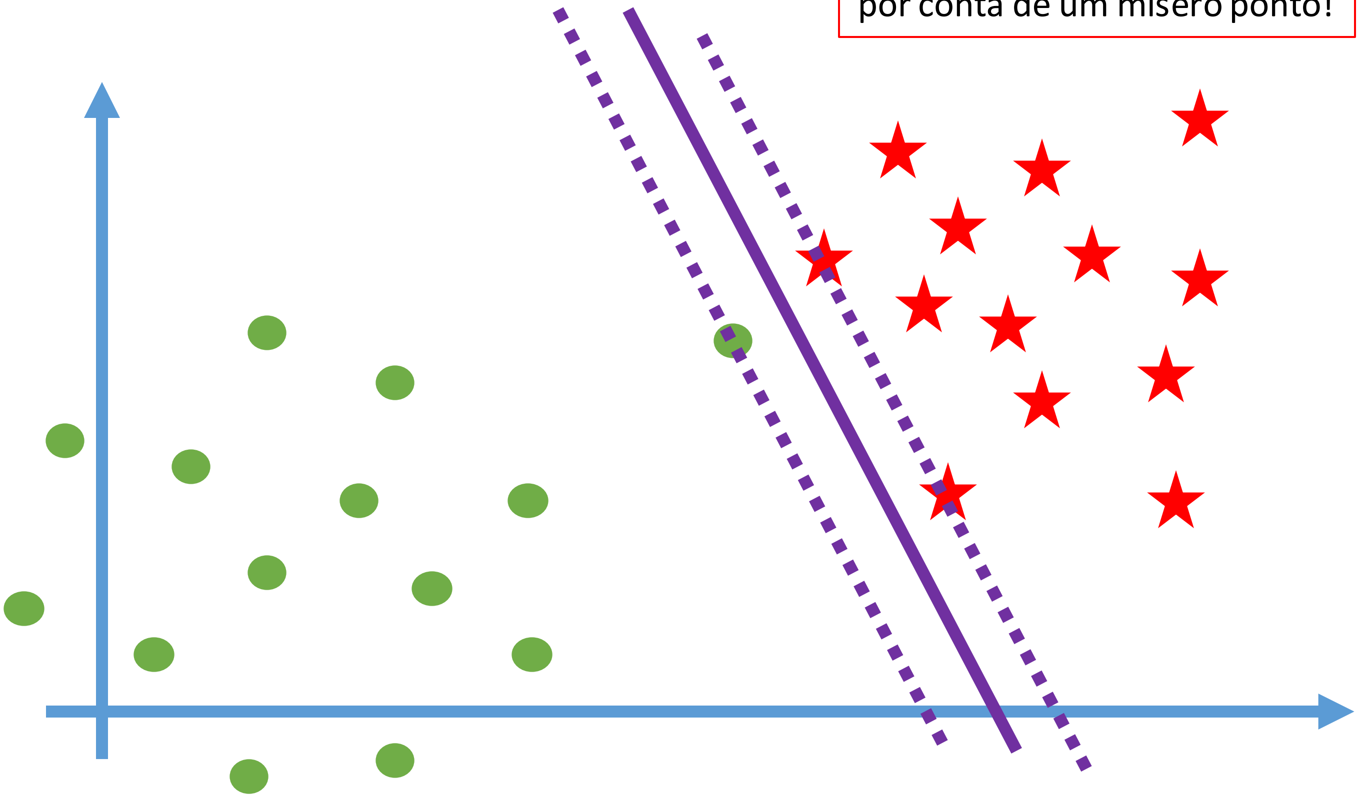


Antes...

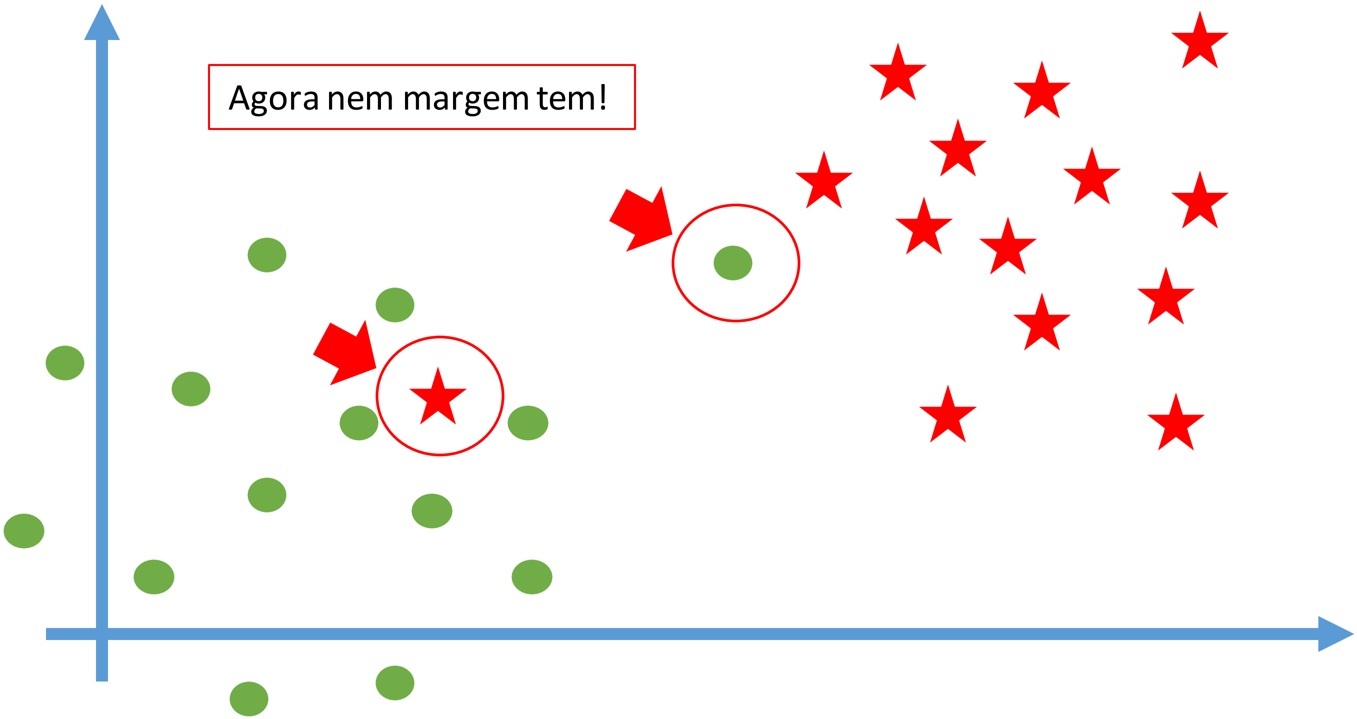


… e depois

Margem muito reduzida
por conta de um mísero ponto!



Mais problemas...



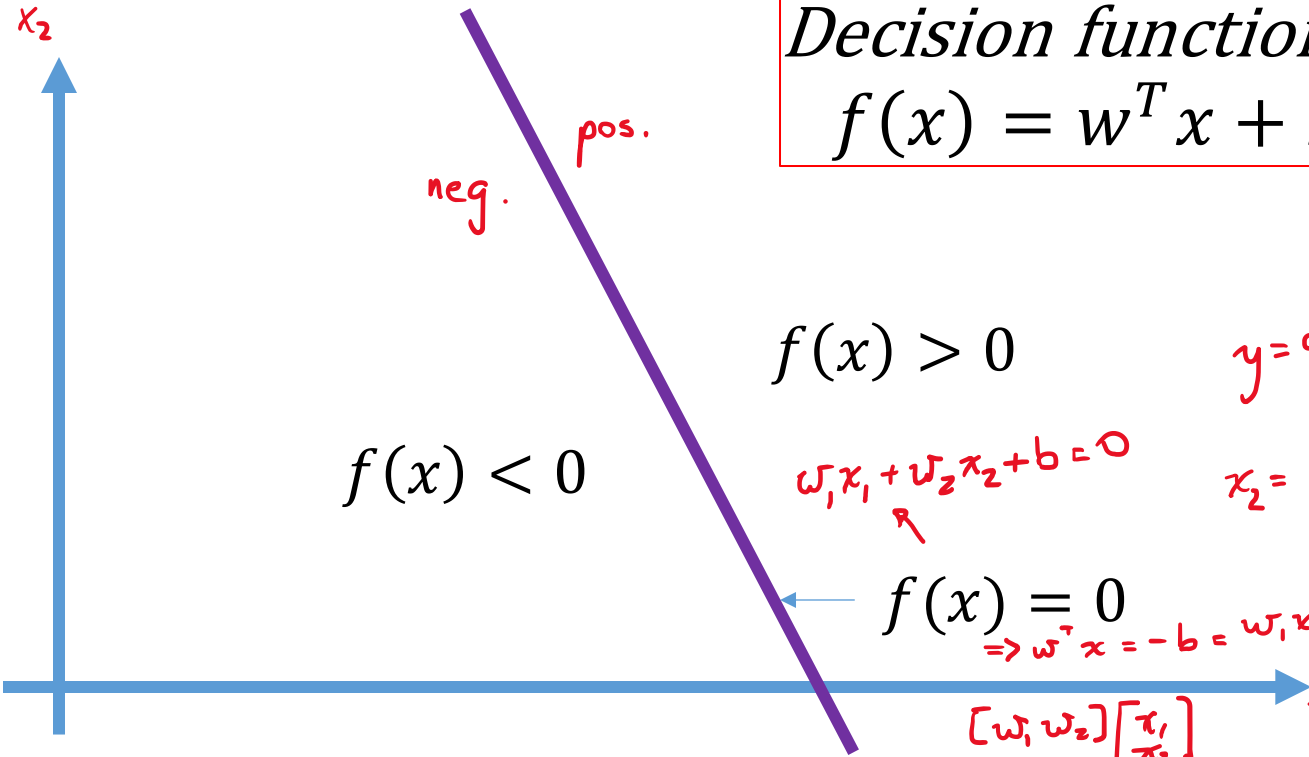
Vamos aos detalhes

- Como formular o problema de “maximizar a avenida”?
- Como lidar com o problema dos outliers?

Equação da reta

Decision function:

$$f(x) = w^T x + b$$



$$y = ax + b$$

$$x_2 = \frac{-w_1 x_1 - b}{w_2}$$

$$w_1 x_1 + w_2 x_2 + b = 0$$

$$\Rightarrow w^T x = -b = w_1 x_1 + w_2 x_2 = -b$$

$$[w_1\ w_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Objetivo

Descobrir qual $f(x)$ implementa a melhor "avenida"

Seja $f(x)$ a melhor função de decisão.

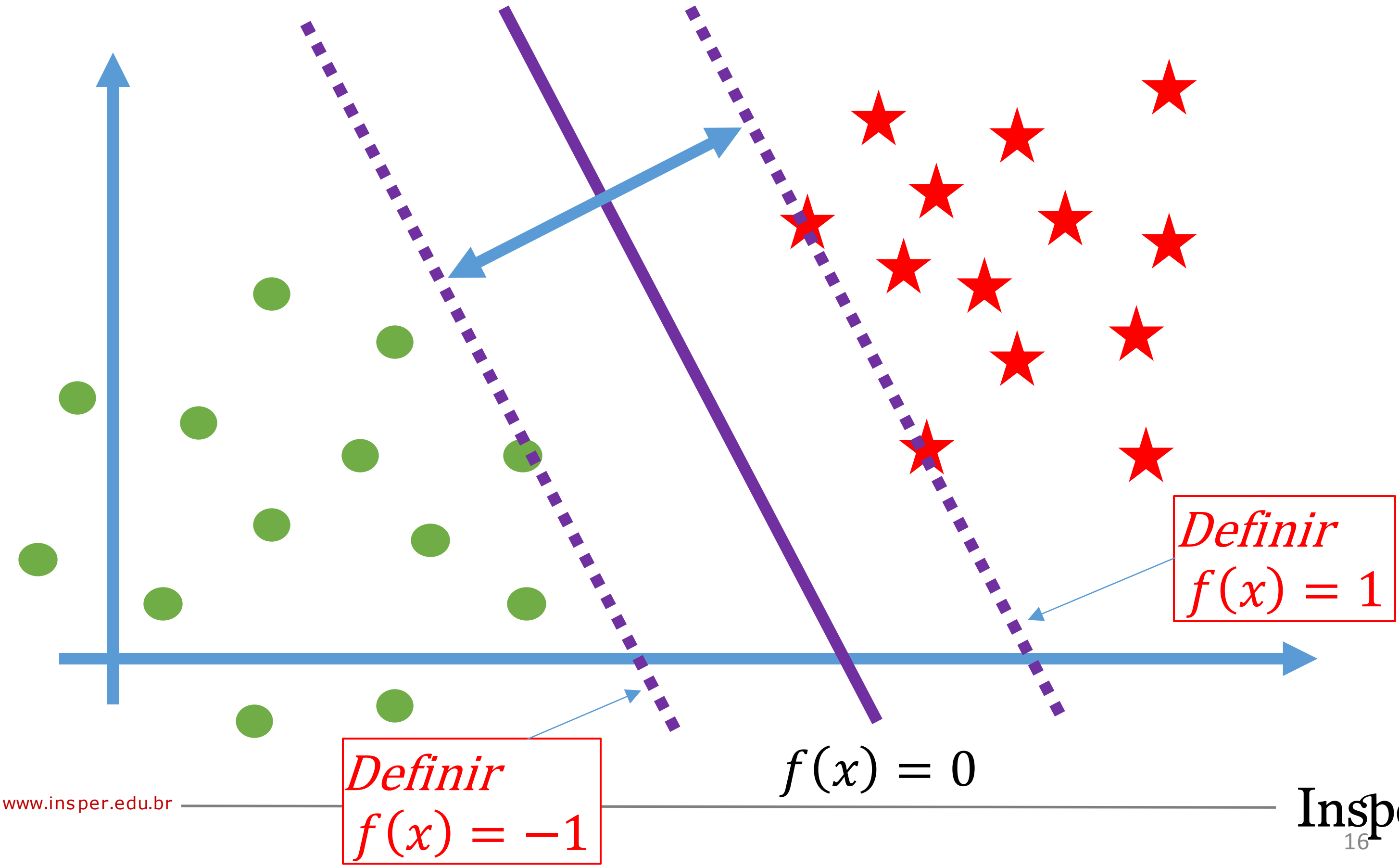
Então $g(x) = K f(x)$ também é igualmente boa!

Afinal, $f(x) = 0 \iff g(x) = 0$

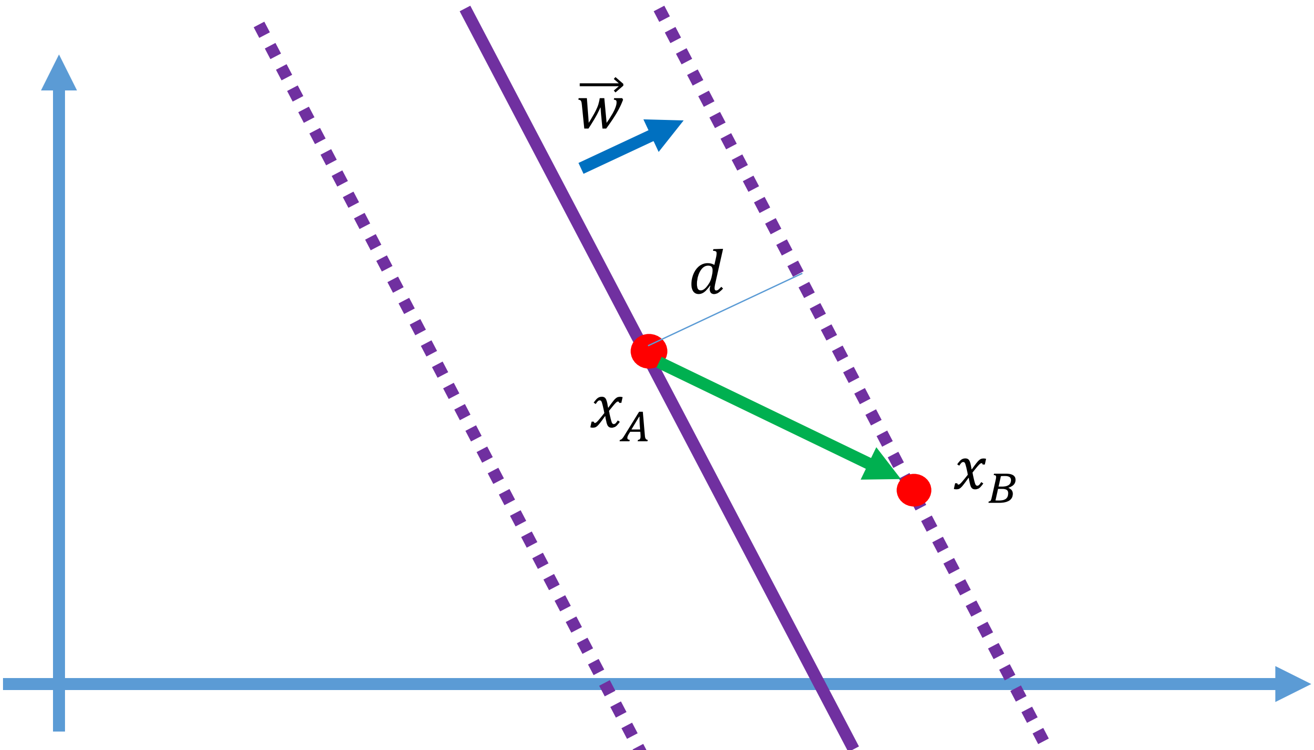
$\Rightarrow$ $f(x)$ e $g(x)$ definem a mesma superfície de separação

AMBIGUIDADE!

Removendo uma ambiguidade...



Tamanho da margem



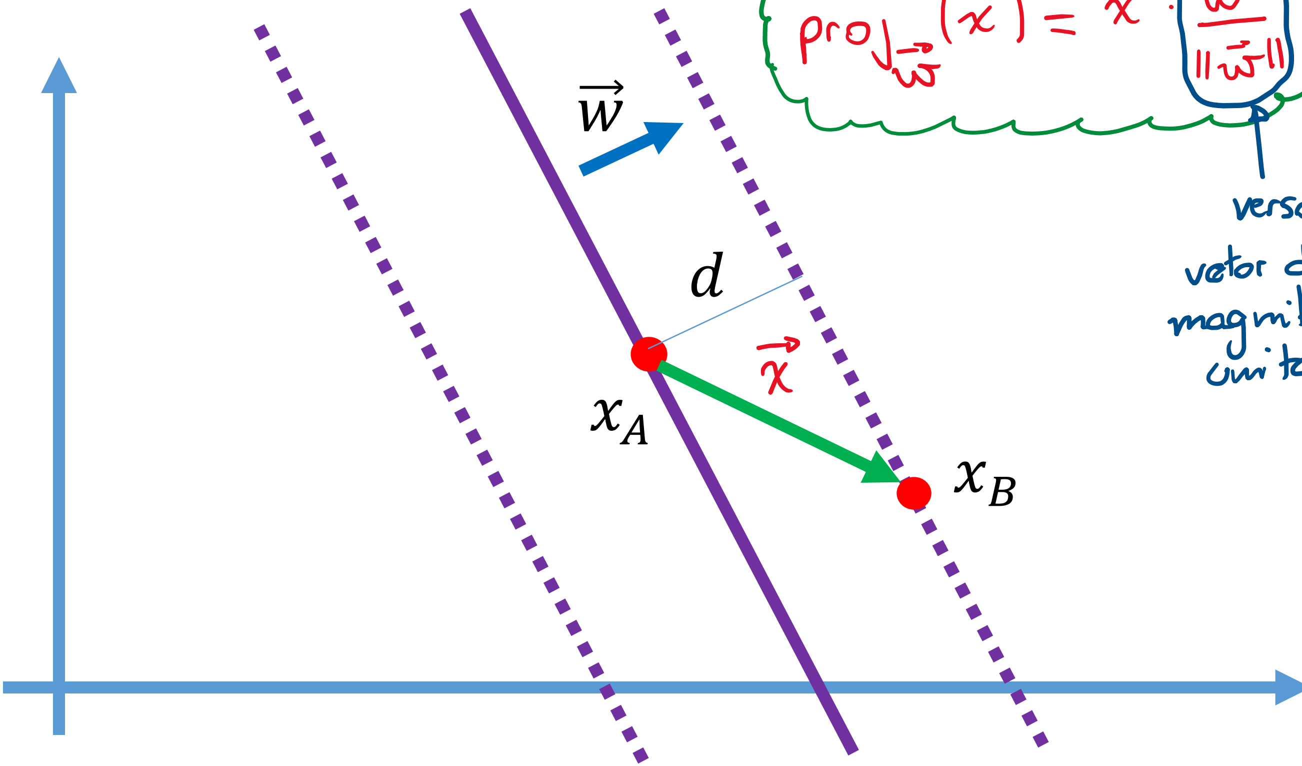
$w^T x + b = 0$

=> dir perpendicular e $\vec{w}$

Tamanho da margem

$$\text{proj}_{\vec{w}}(\vec{x}) = \vec{x} \cdot \boxed{\frac{\vec{w}}{\|\vec{w}\|}}$$

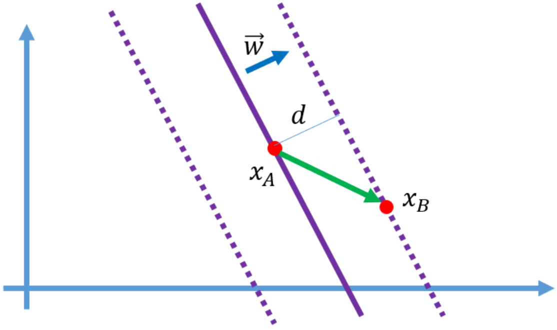
versor: vetor de magnitude unitária



Tamanho da margem

Em A: $f(x_A) = 0 \Rightarrow w^T x_A + b = 0 \Rightarrow w^T x_A = -b$

Em B: $f(x_B) = 1 \Rightarrow w^T x_B + b = 1 \Rightarrow w^T x_B = 1 - b$



Tamanho da semi-margem: projeção na direção $\vec{w}$

$$d = \frac{(\overrightarrow{x_B} - \overrightarrow{x_A}) \cdot \vec{w}}{\|\vec{w}\|} = \frac{w^T(x_B - x_A)}{\sqrt{w^T w}}$$

Portanto:

$$d = \frac{w^T x_B - w^T x_A}{\sqrt{w^T w}} = \frac{1 - b - (-b)}{\sqrt{w^T w}} = \frac{1}{\sqrt{w^T w}}$$

$\sum_{i=1}^{n} w_i^2$

Ou seja: maximizar d equivale a minimizar $\overbrace{w^T w}$

Problema de otimização da SVM

minimizar $w^T w$

sujeito a: respeitar a "regra da calçada"

— " —

minimizar $w^T w$

sujeito a: $f(x_i) \geq 1$ p/ "amostras positivas"

$f(x_i) \leq -1$ p/ "amostras negativas"

$\Downarrow$

$w^T x_i + b \geq 1$ se $y_i = 1$

$w^T x_i + b \leq -1$ se $y_i = -1$

Em SVM, dizemos que (-1) indica classe negativa.

$w^T x_i + b \geq 1$ se $y_i = 1 \Rightarrow y_i(w^T x_i + b) \geq 1$ se $y_i = 1$

$w^T x_i + b \leq -1$ se $y_i = -1 \Rightarrow y_i(w^T x_i + b) \geq 1$ se $y_i = -1$

$$\Rightarrow \boxed{y_i(w^T x_i + b) \geq 1, \ \forall i}$$

otimização quadrática

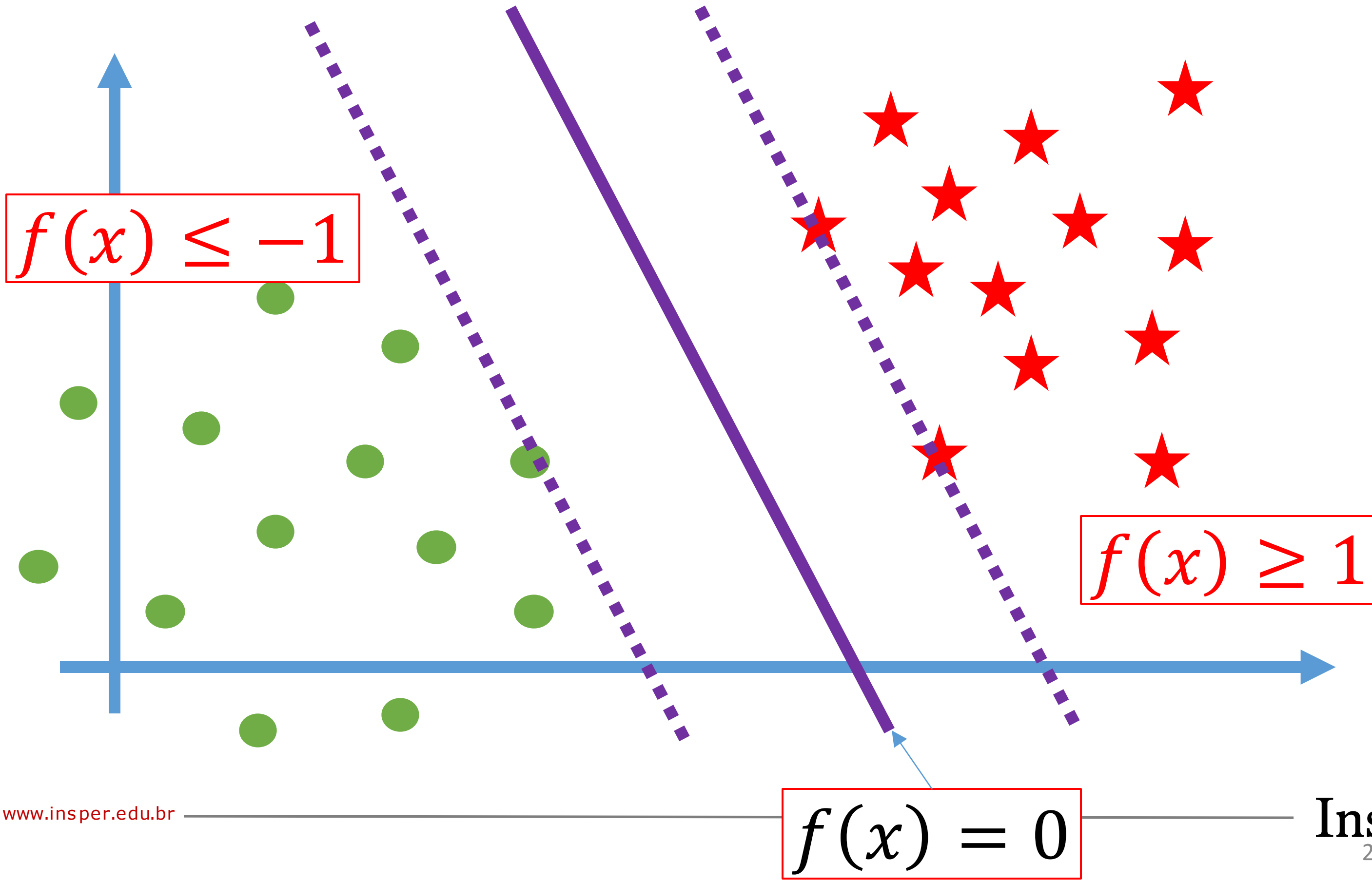
<u>SVM</u>: (hard margin)

minimizar $w^T w$

$\Rightarrow$ sujeito a: $y_i(w^T x_i + b) \geq 1, \ \forall i$

onde vale a igualdade
$\Rightarrow$ vetores de suporte.

Critério: pontos fora da “avenida”



Critério: pontos fora da "avenida"

Truque: defina $t_i = \begin{cases} 1\ se\ x_i\ cai\ do\ lado\ f(x) > 0 \\ -1\ se\ x_i\ cai\ do\ lado\ f(x) < 0 \end{cases}$

Vamos pensar um pouco: o que acontece com os valores $t_i f(x_i)$ se o critério de "pontos fora da avenida" é respeitado?

Support Vector Machines

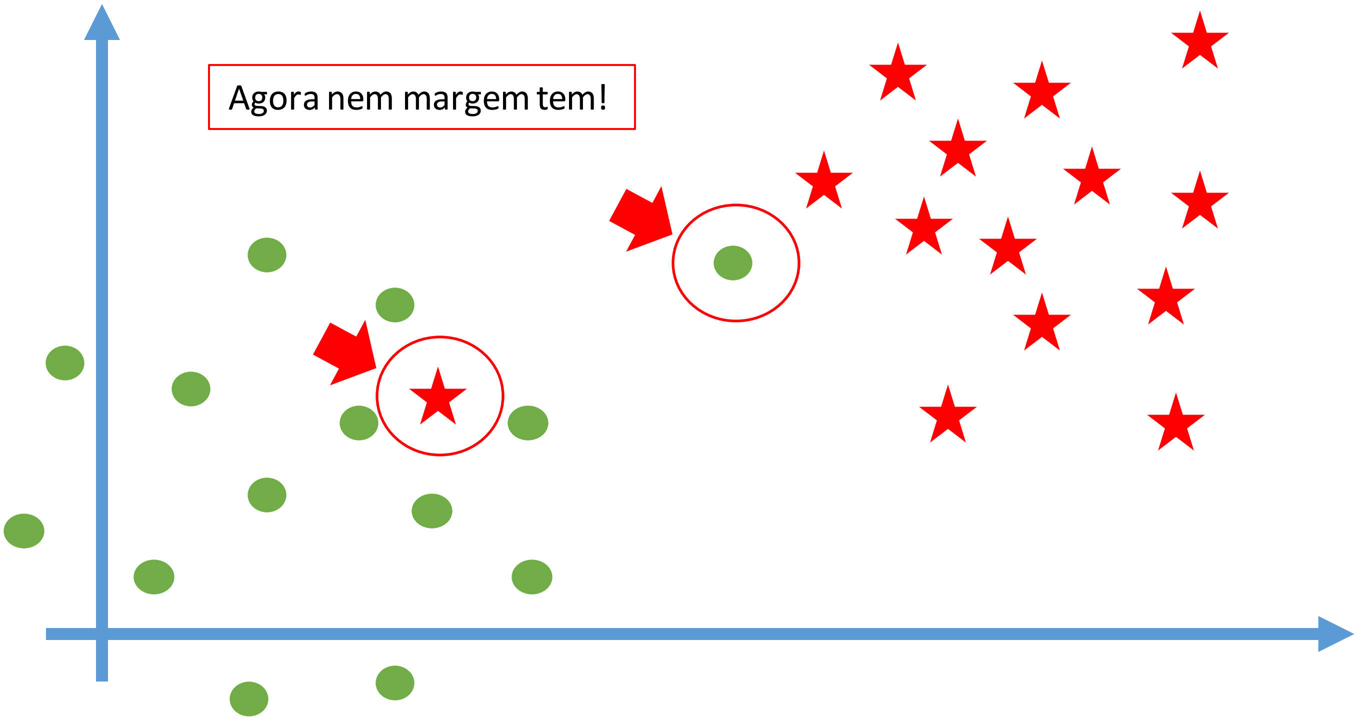
minimizar $\frac{1}{2} w^T w$

Maximizar a margem de classificação

sujeito a $t_i(w^T x_i - b) \geq 1$,
para $i = 1, 2, \cdots, m$

Respeitar o critério de
"pontos fora da margem"

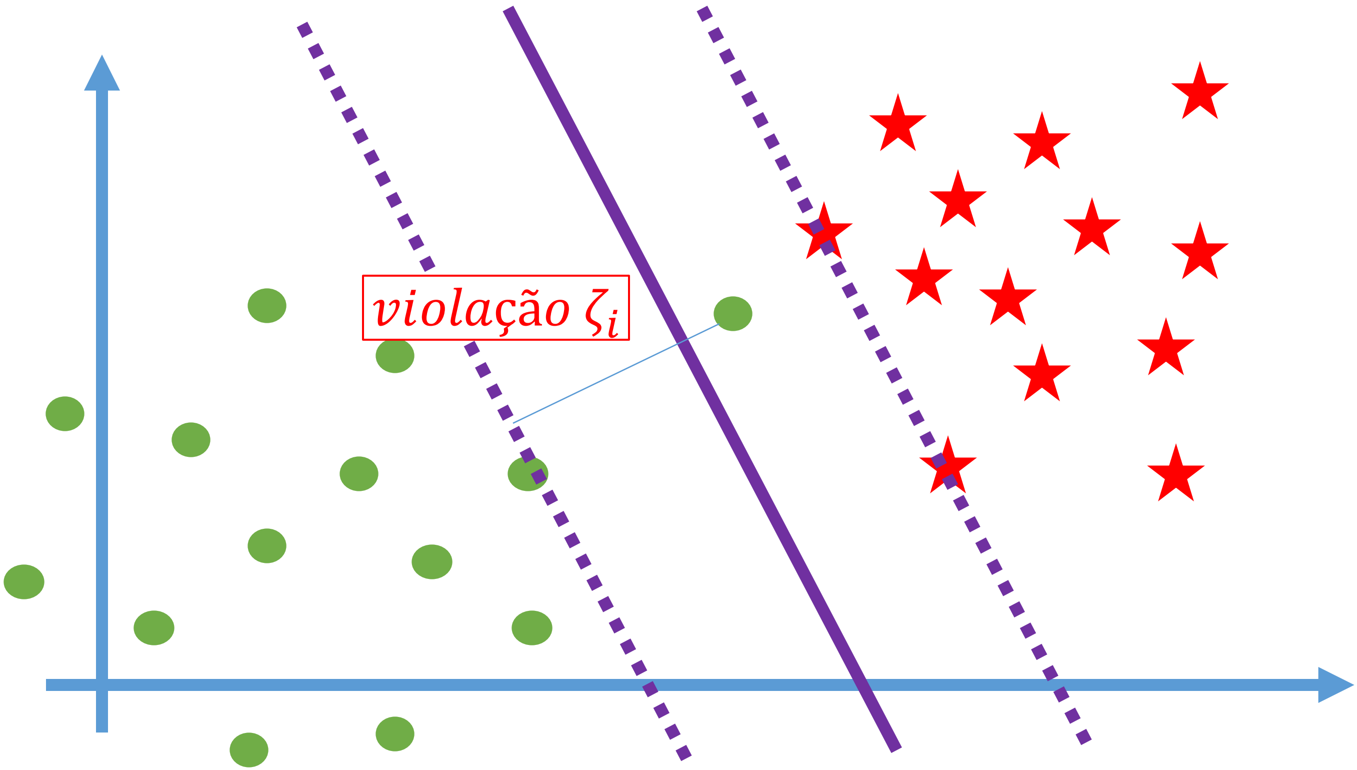
E como fica esse caso?



Vamos pensar um pouco

- Se um ponto viola a condição de “pontos fora da margem”, o que acontece de errado na formulação matemática da SVM?

Pedágio da SVM...



Pedágio da SVM...

- Ok, vamos aceitar violações ζ_i mas a um custo $C\zeta_i$

- Pontos que não violam o critério da SVM terão violação $\zeta_i = 0$, e portanto não pagam a penalidade.

SVM, soft-margin

$$\text{minimizar } \frac{1}{2}w^T w + C\sum_{i=1}^{m} \zeta_i$$

Maximizar a margem de classificação
com penalidade

Hiperparâmetro!

pontos devem estar fora da avenida...

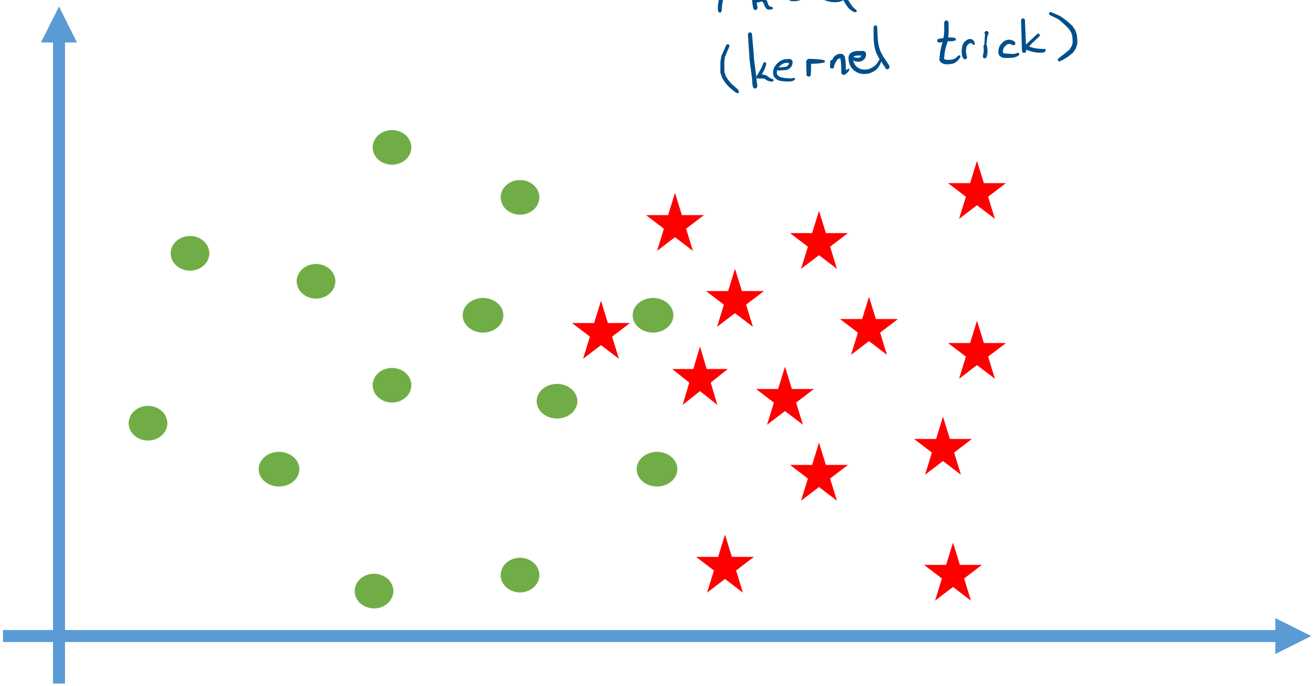
ou pagar penalidade

$$\text{sujeito a } \underbrace{t_i(w^T x_i - b) \geq (1}_{} - \zeta_i) \text{ e } \zeta_i \geq 0$$
$$\text{para } i = 1, 2, \cdots, m$$

Respeitar o critério de
"pontos fora da margem"
com permissão de outliers

Um problema de classificação

TRUQUE DO KERNEL
(kernel trick)



Kernel

$K(x_1, x_2)$: "tipo um produto escalar"

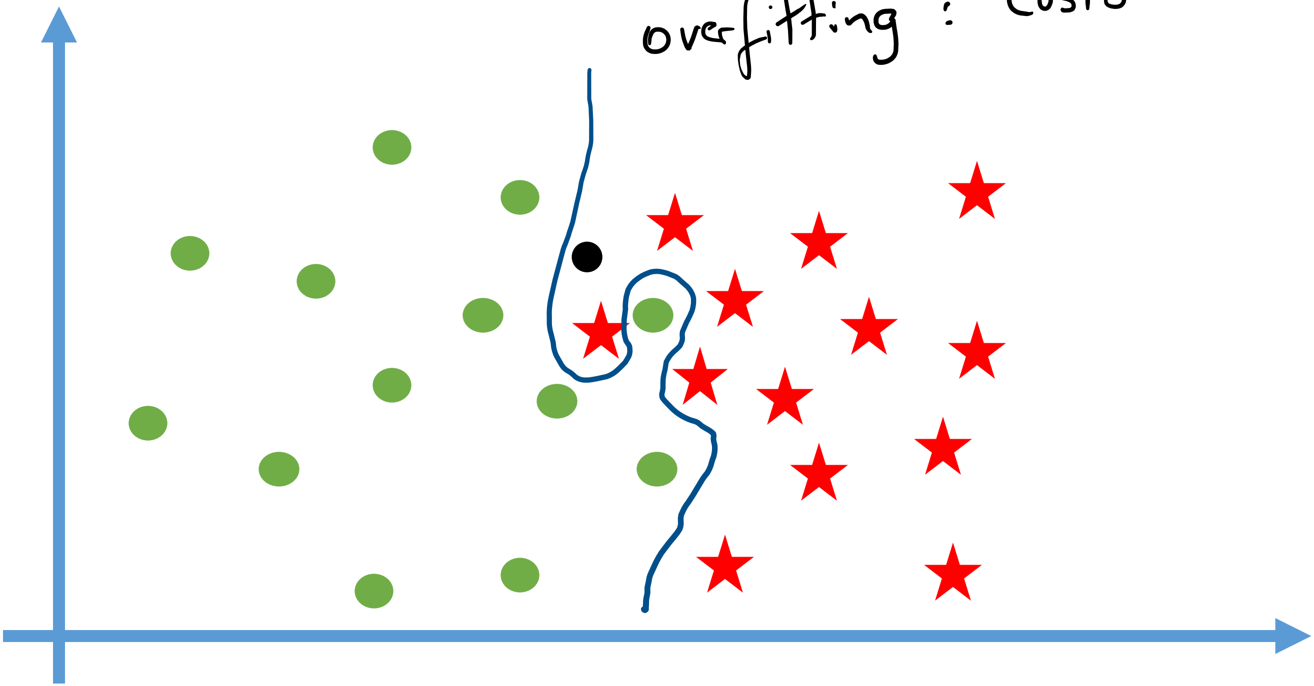
- $K(x_1, x_1) \geq 0$
- $c^T K(x, y) c \geq 0$, $\forall x, y, c$ (?)

$$K(x_1, x_2) = \phi^T(x_1)\phi(x_2)$$

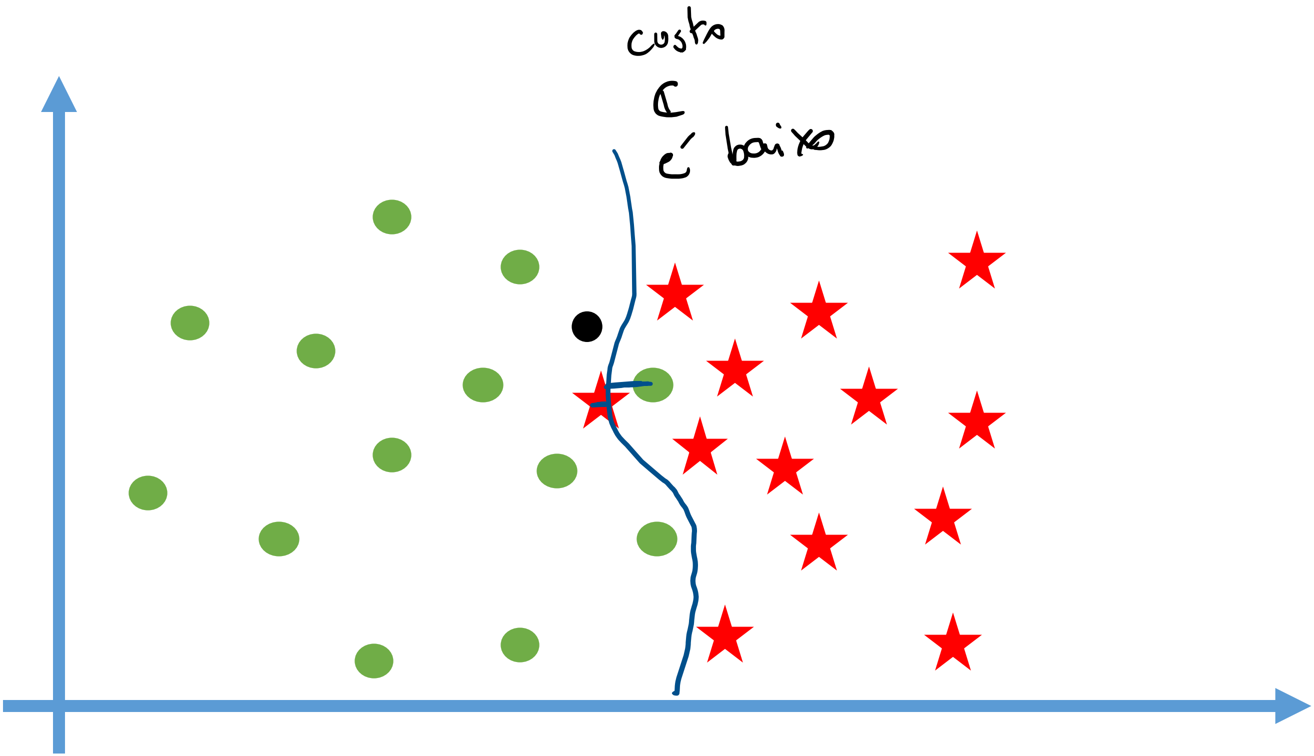
Teorema de Mercer

Um problema de classificação

overfitting : custo C é alto

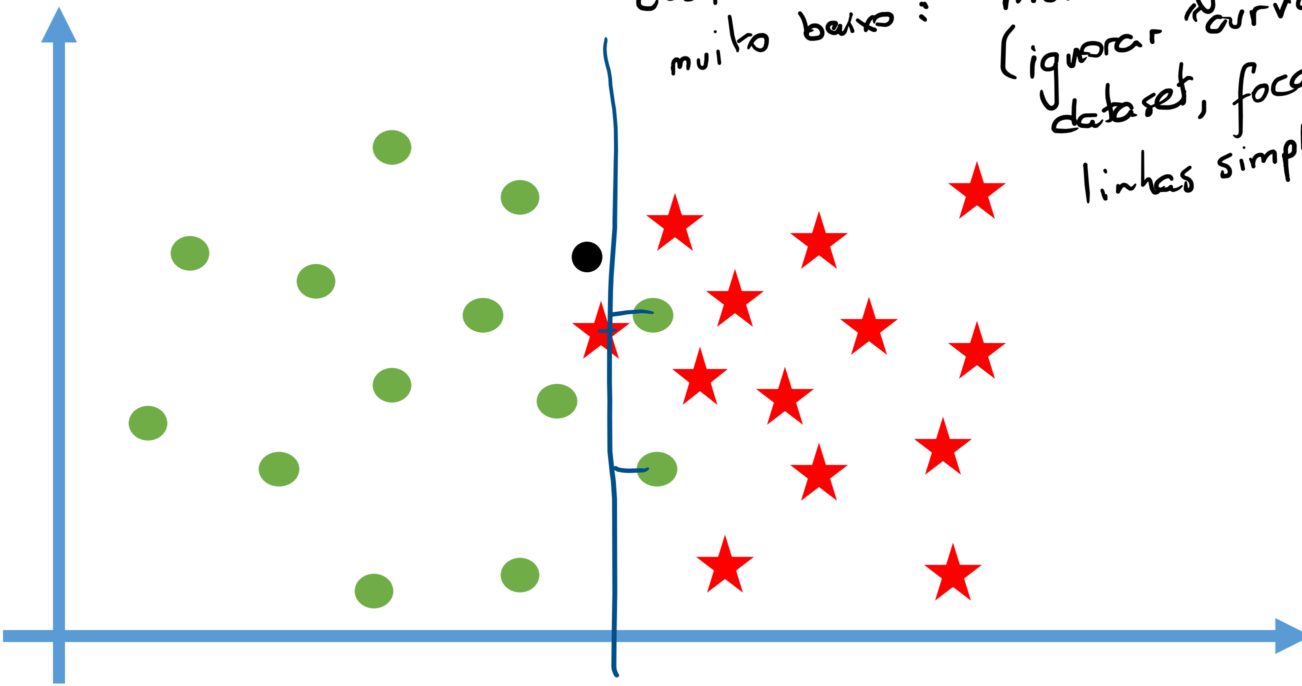


Um problema de classificação

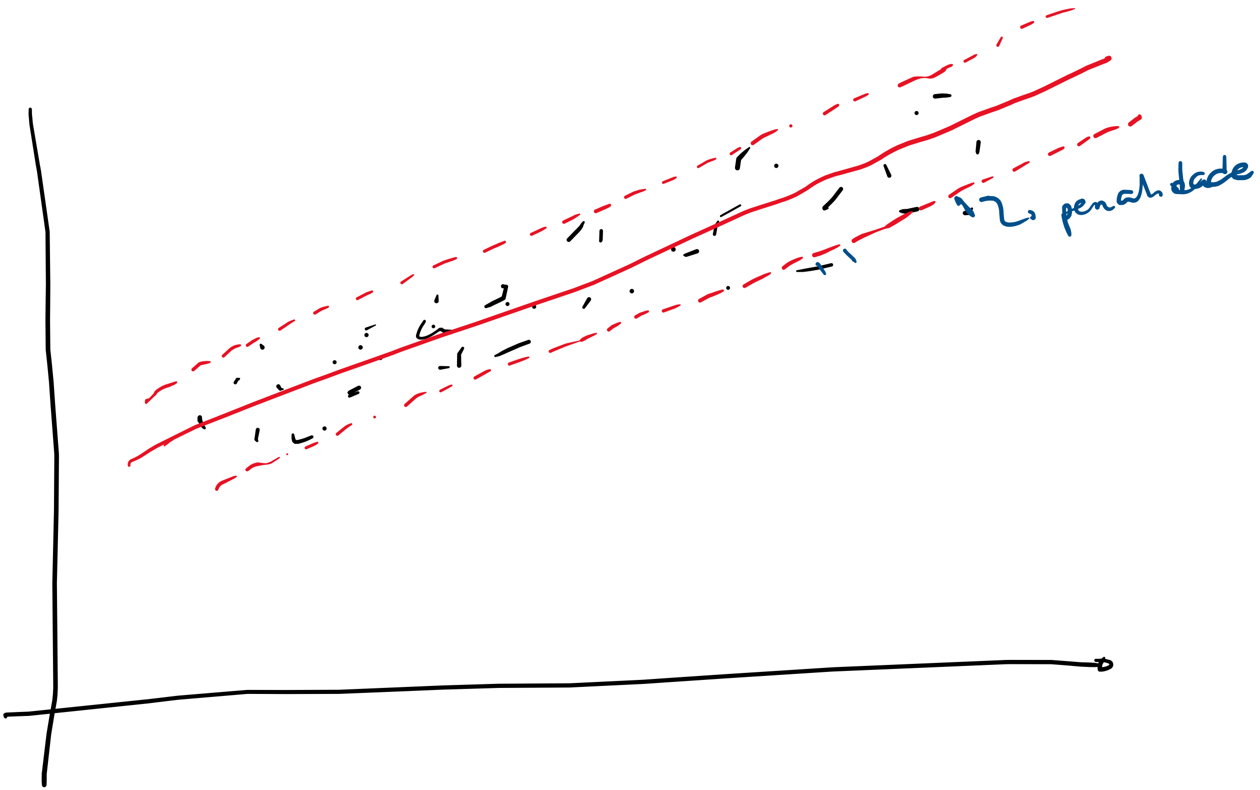


Um problema de classificação

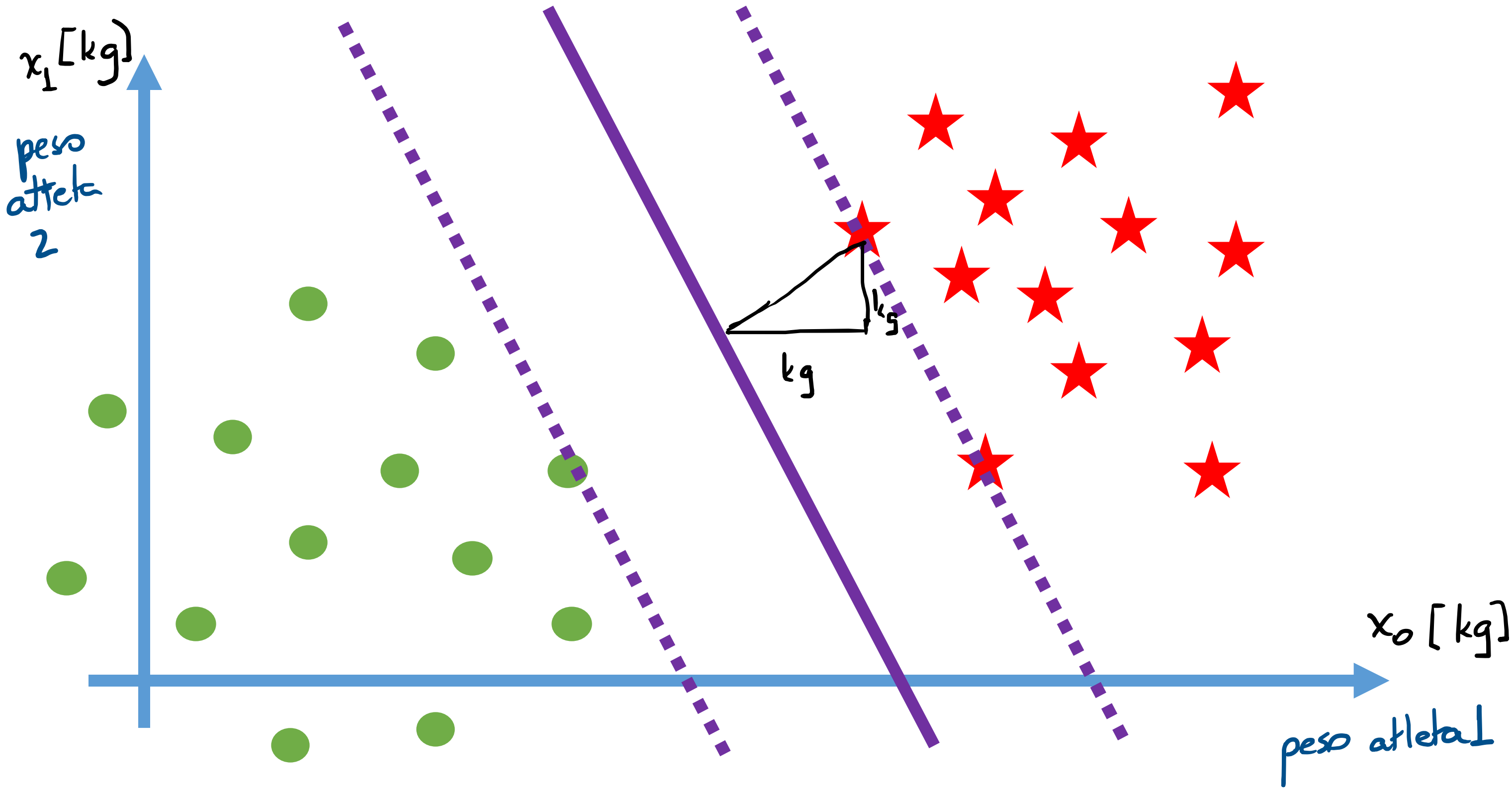
custo C é muito baixo: muita regularização (ignorar "curvas" do dataset, focar em linhas simples)



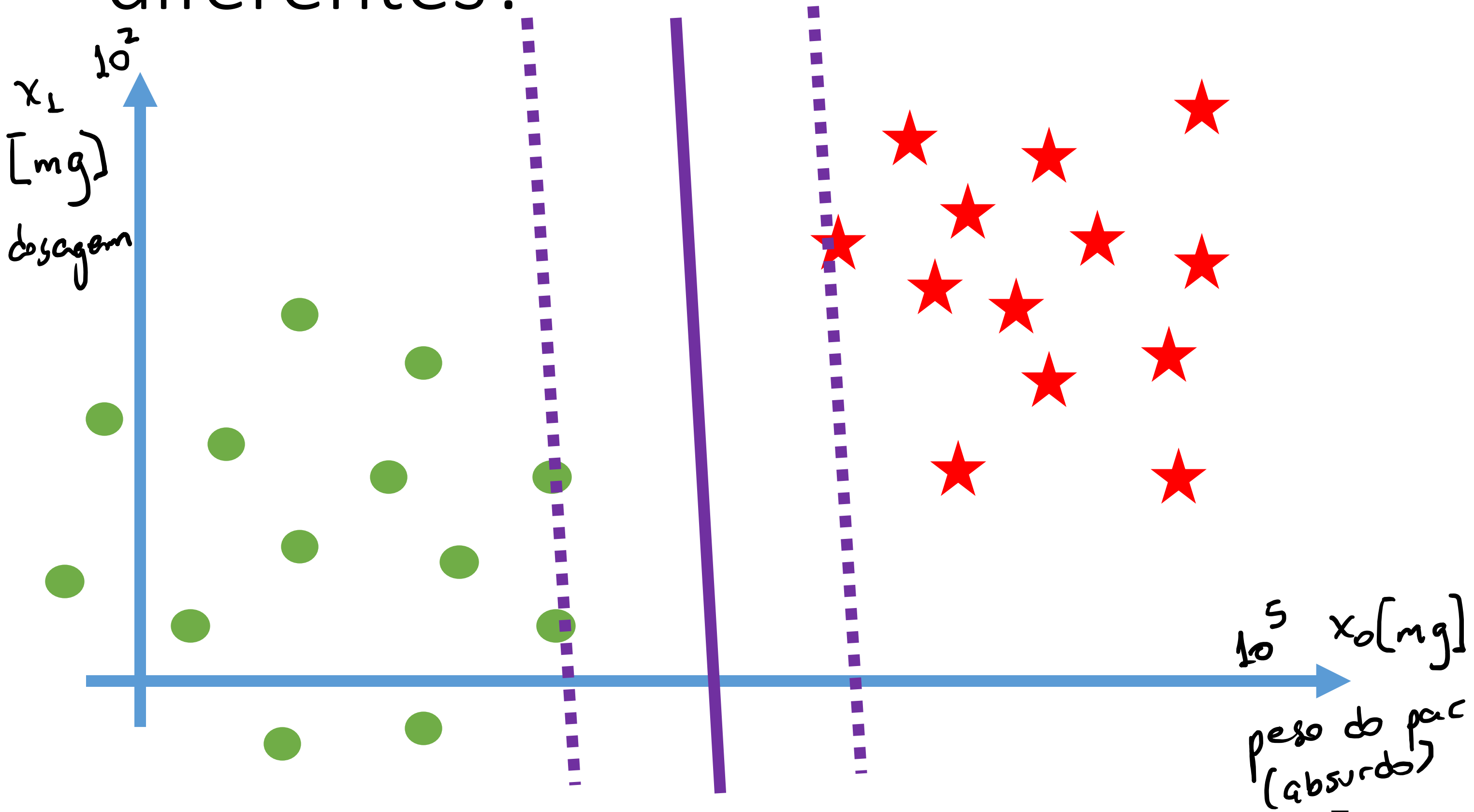
Regressão



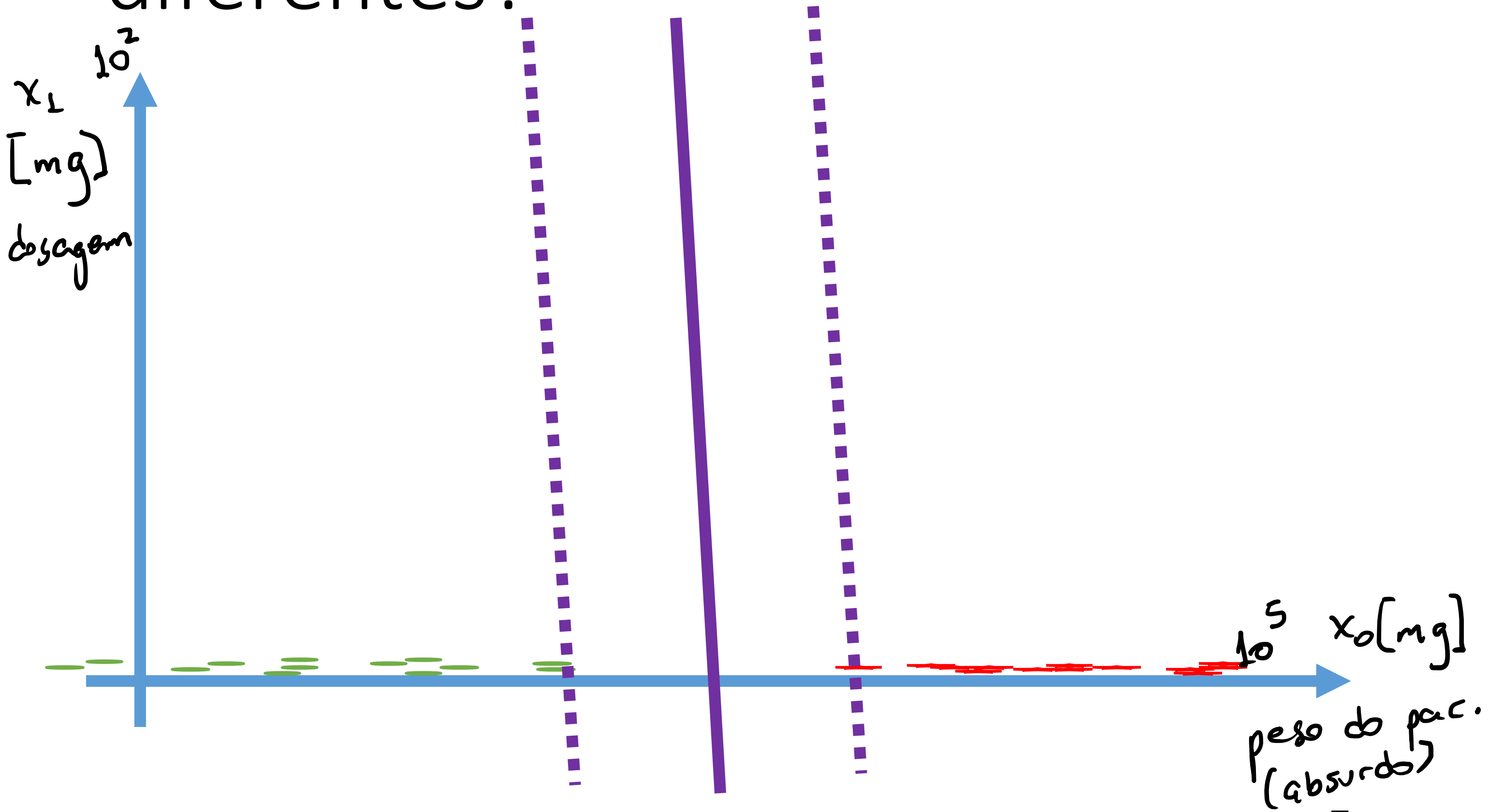
E se os eixos tem escalas diferentes?



E se os eixos tem escalas diferentes?



E se os eixos tem escalas diferentes?



Conclusão

SVM é MUITO sensível a escala dos eixos

=> Quase obrigatório StandardScaler()

Complexidade do Support Vector Machines no treinamento

Linear : $O(mn)$

- Se m dobra, o tempo de treina/o dobra.
- Se n dobra, " " " ".

Não-linear (kernels) : $O(m^2 n)$ a $O(m^3 n)$ (depende do kernel)

- Se m dobra, o tempo de treina/o $\times 4$ ou $\times 8$.
- Se n dobra, o tempo de treina/o dobra.

Insper