



Insper

Machine Learning

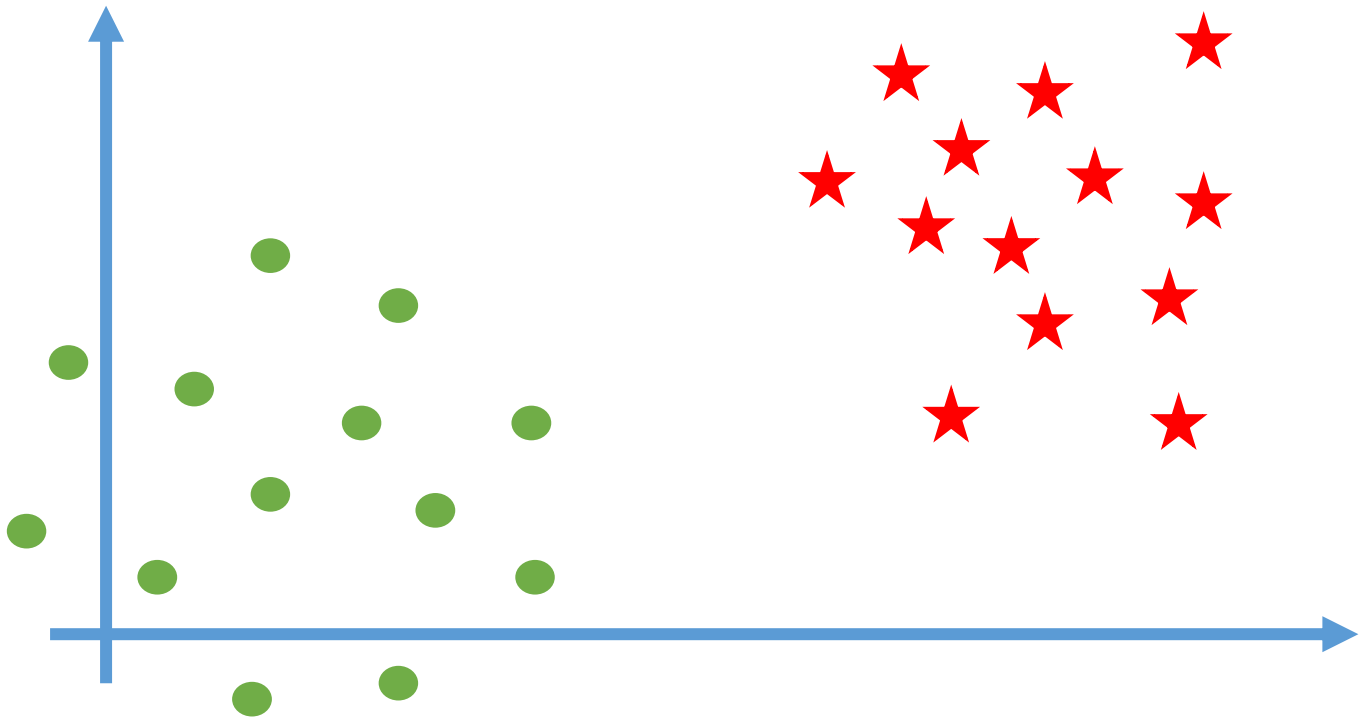
Aula 11 – Support Vector Machines

2021 – Engenharia
Fábio Ayres <fabioja@insper.edu.br>

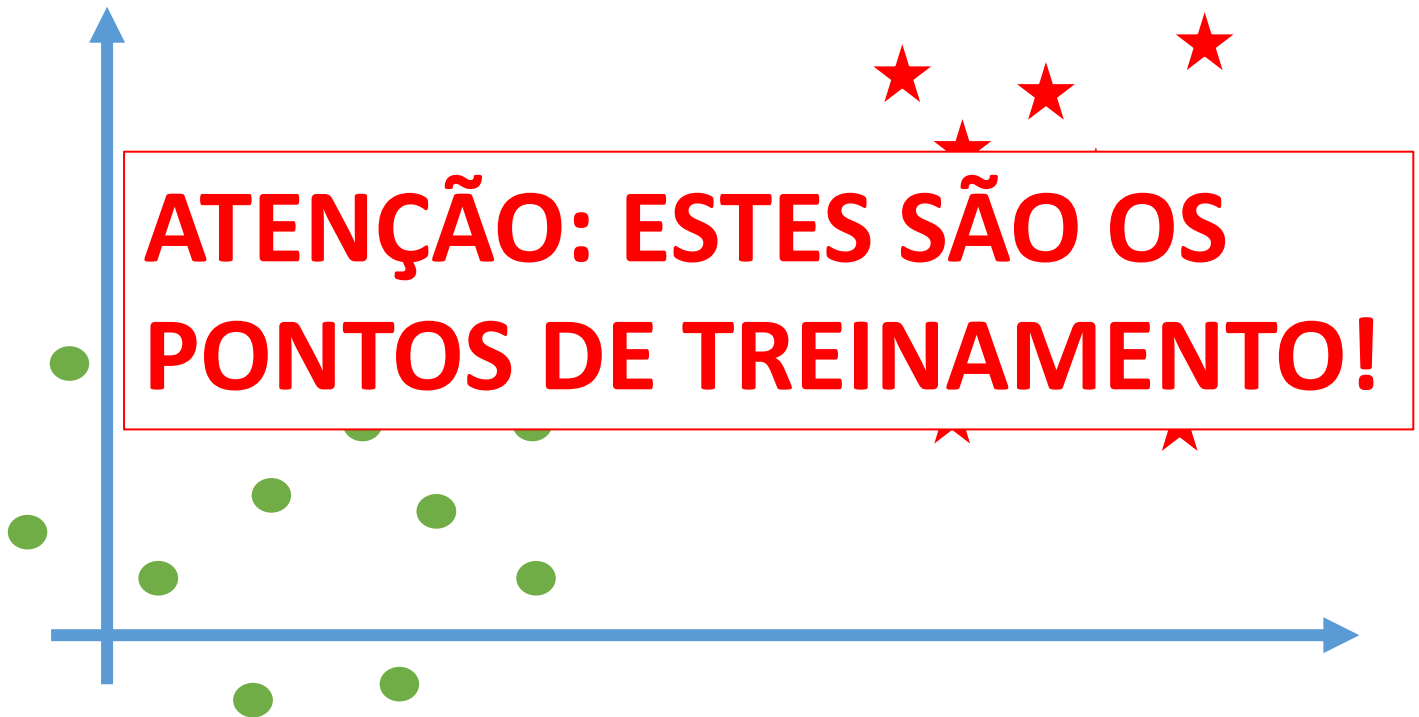
Objetivos da aula

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

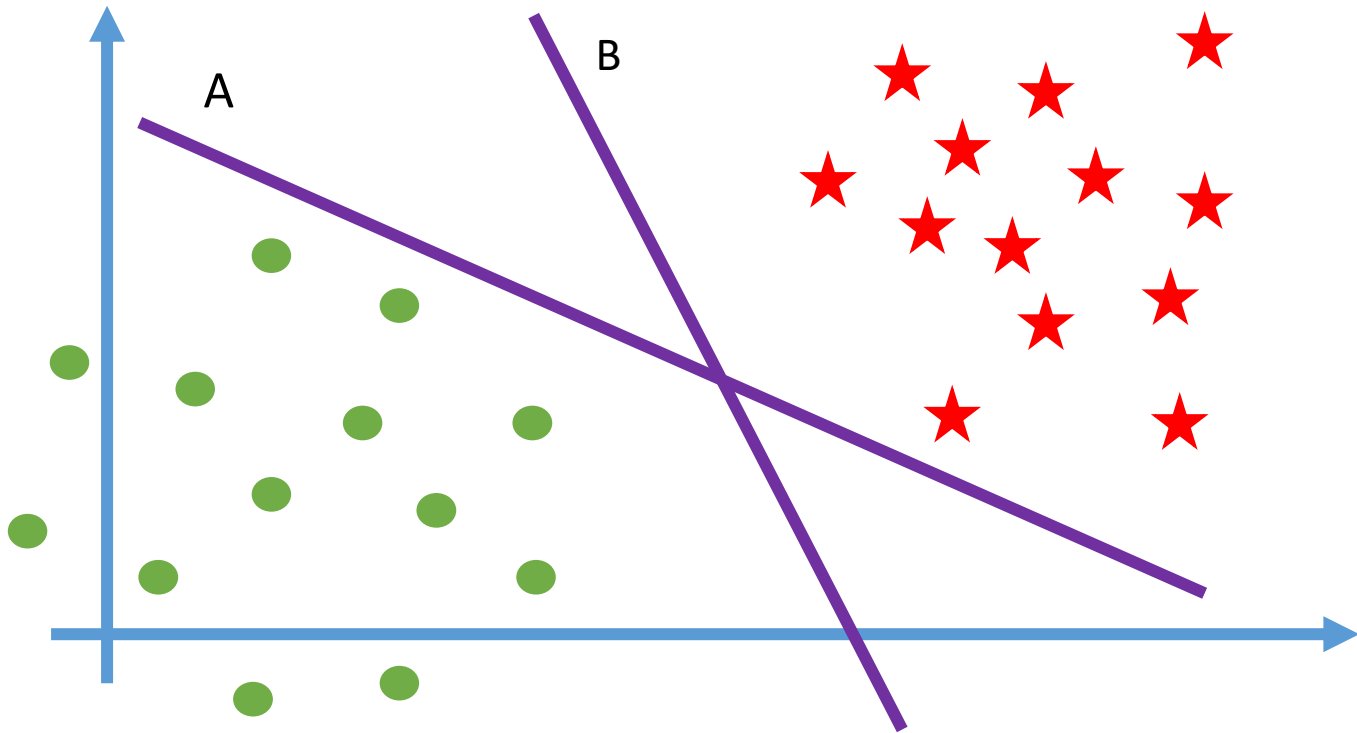
Um problema de classificação



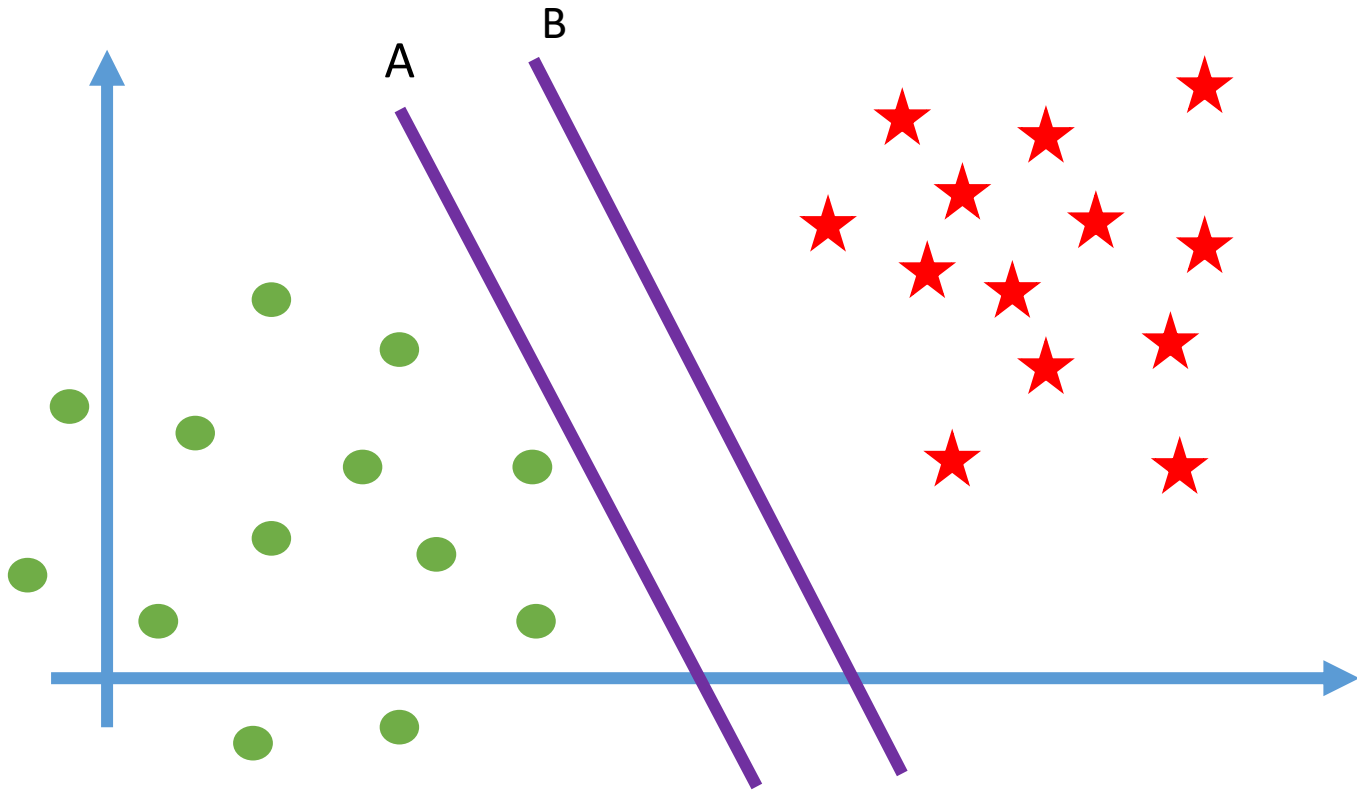
Um problema de classificação



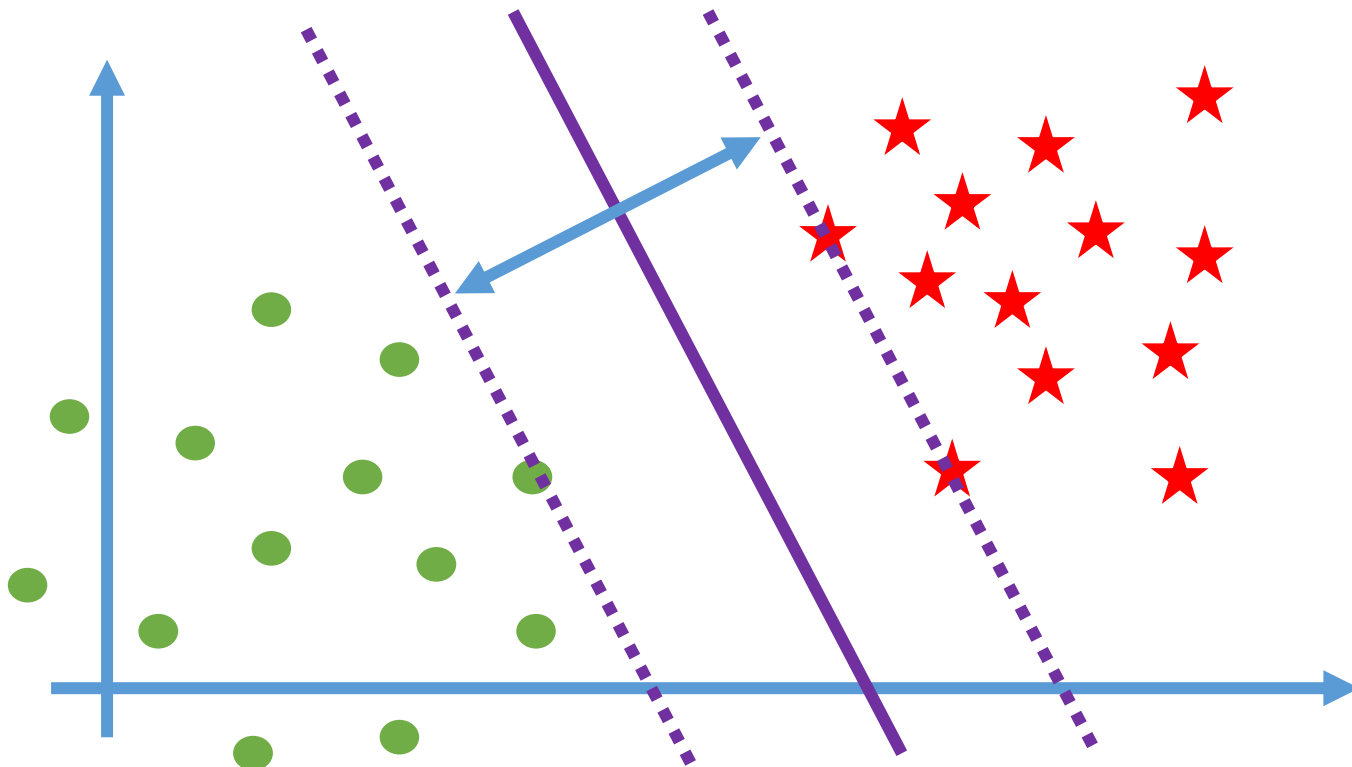
Qual a melhor reta de separação?



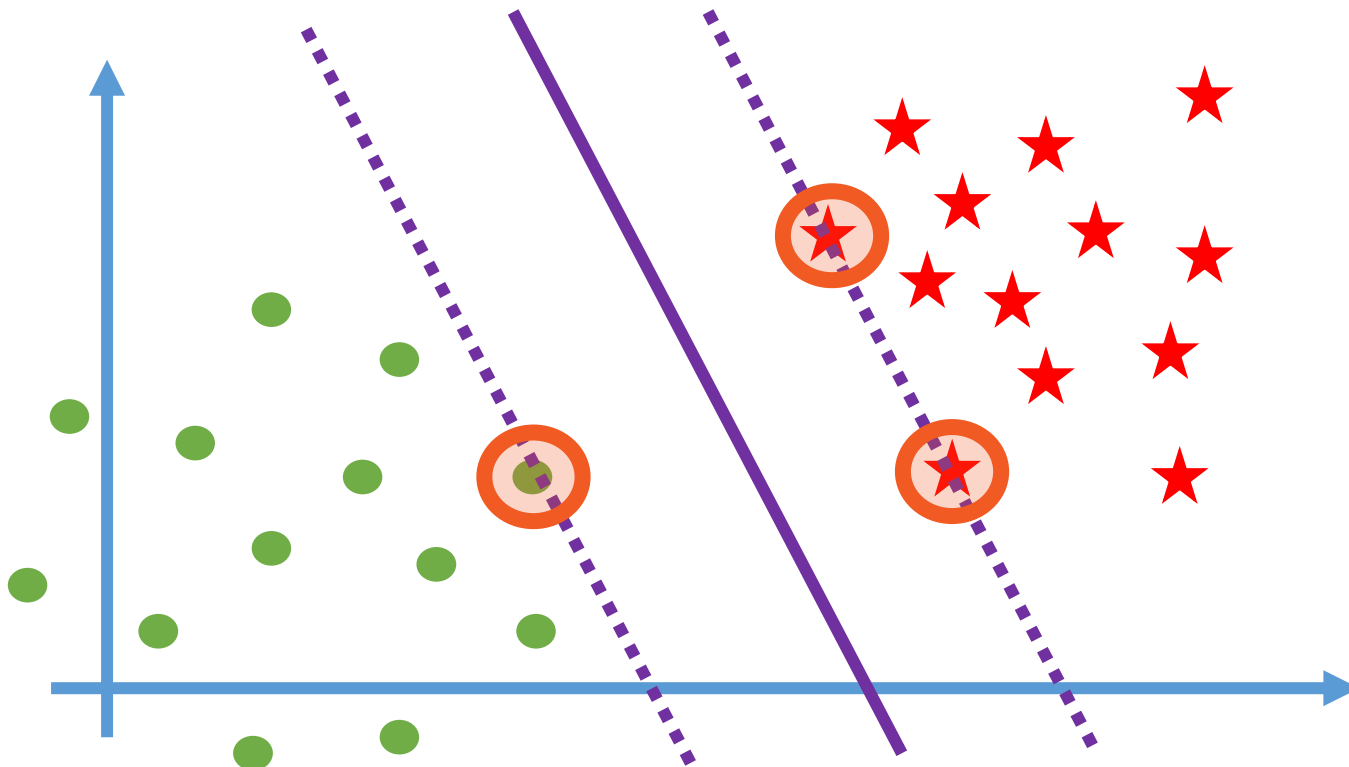
Qual a melhor reta de separação?



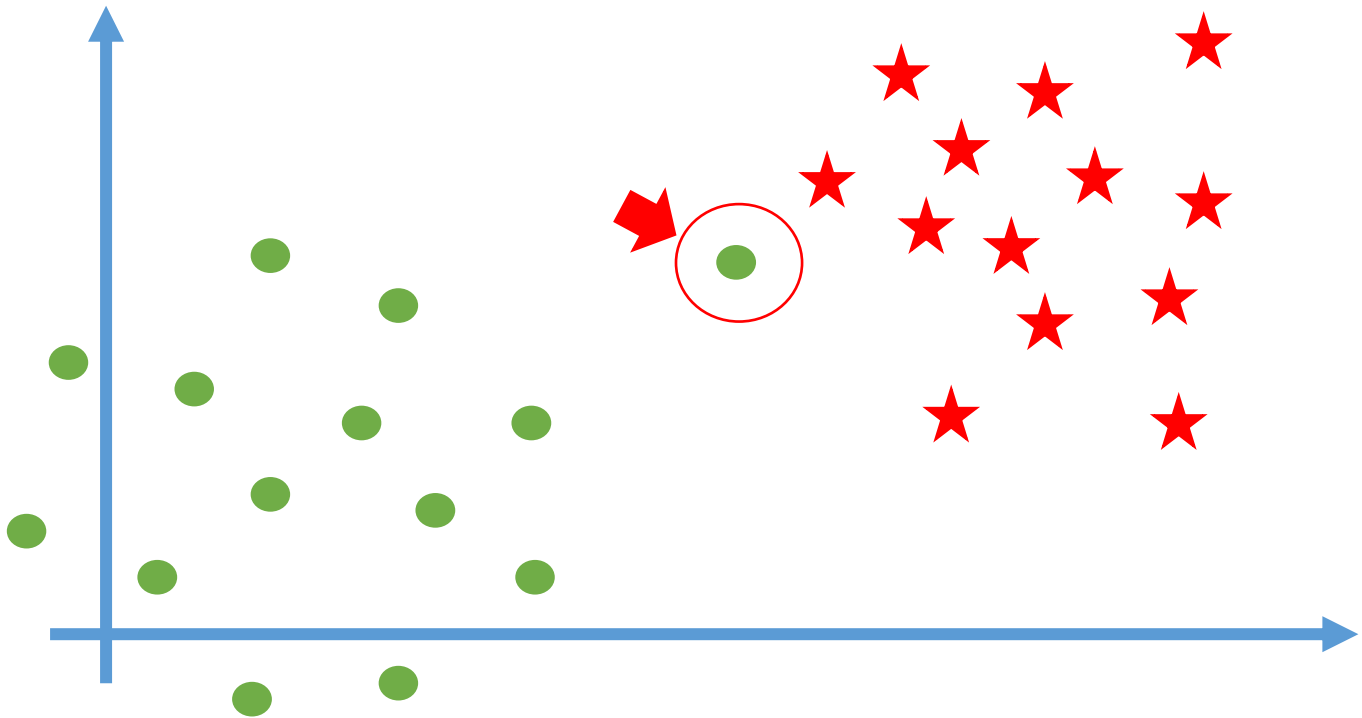
Ideia: aumentar a “avenida”



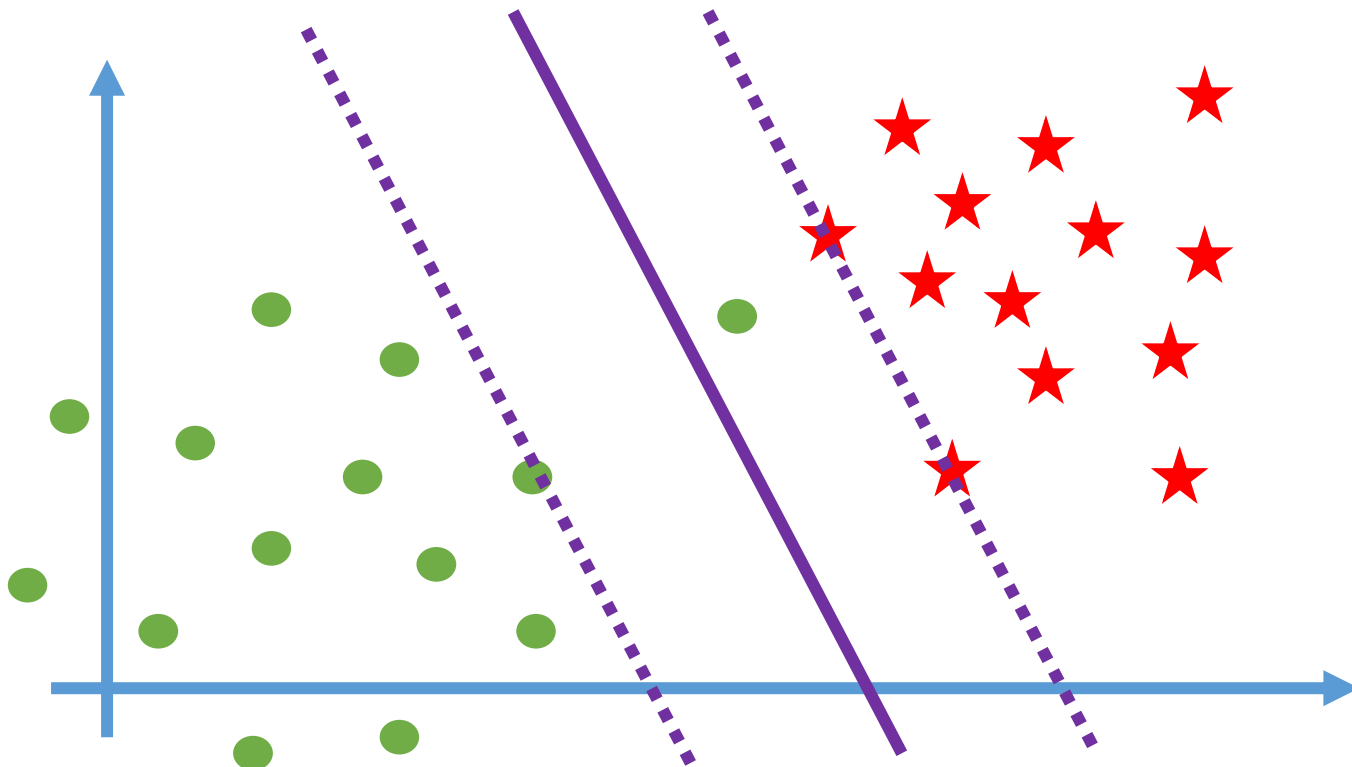
Vetores de suporte



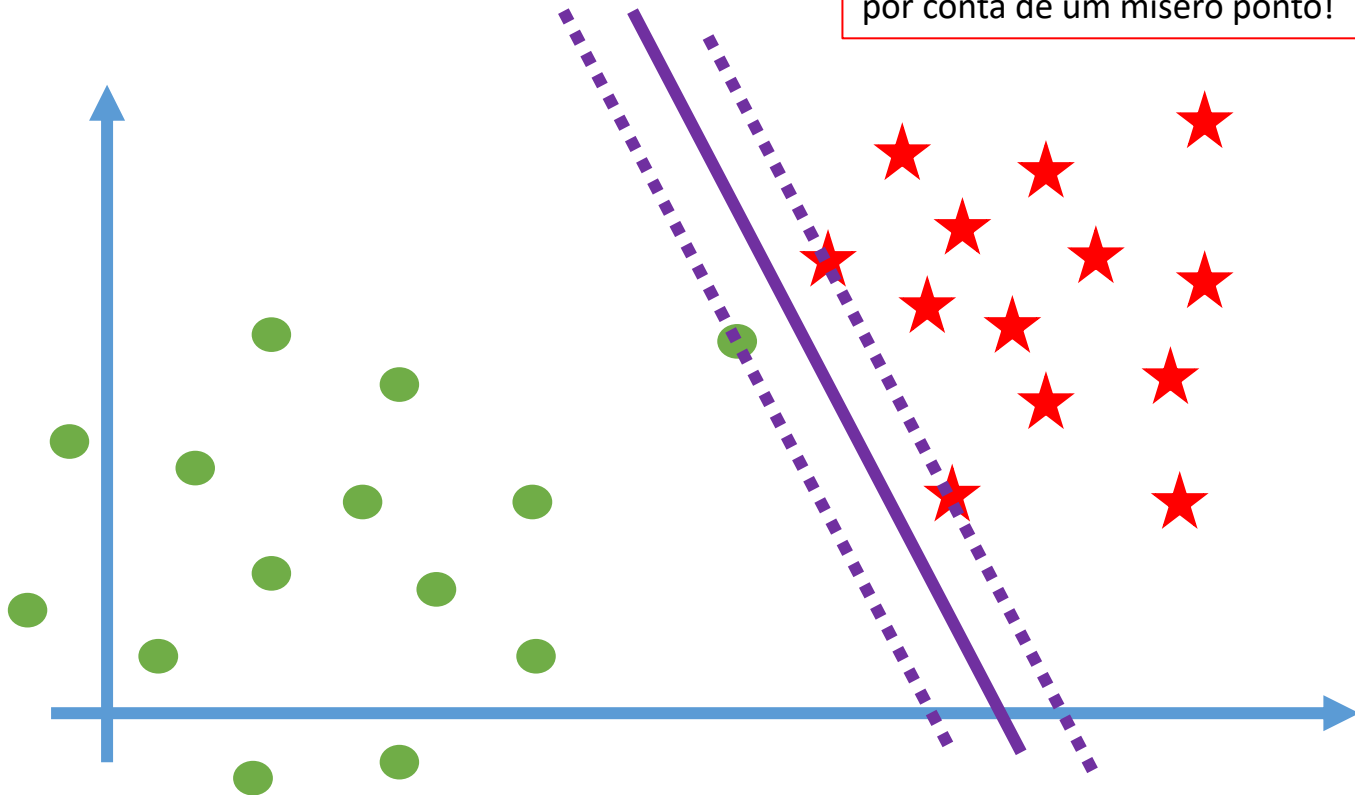
Problemas no paraíso...



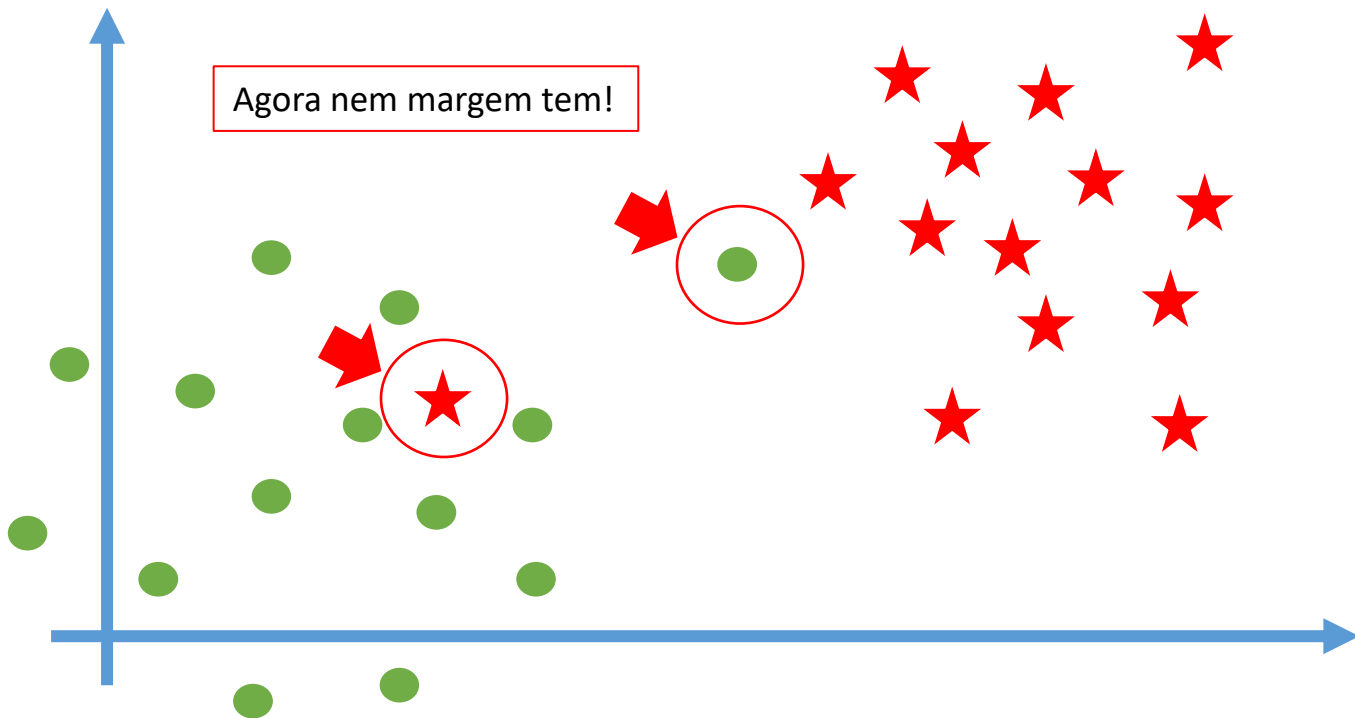
Antes...



... e depois



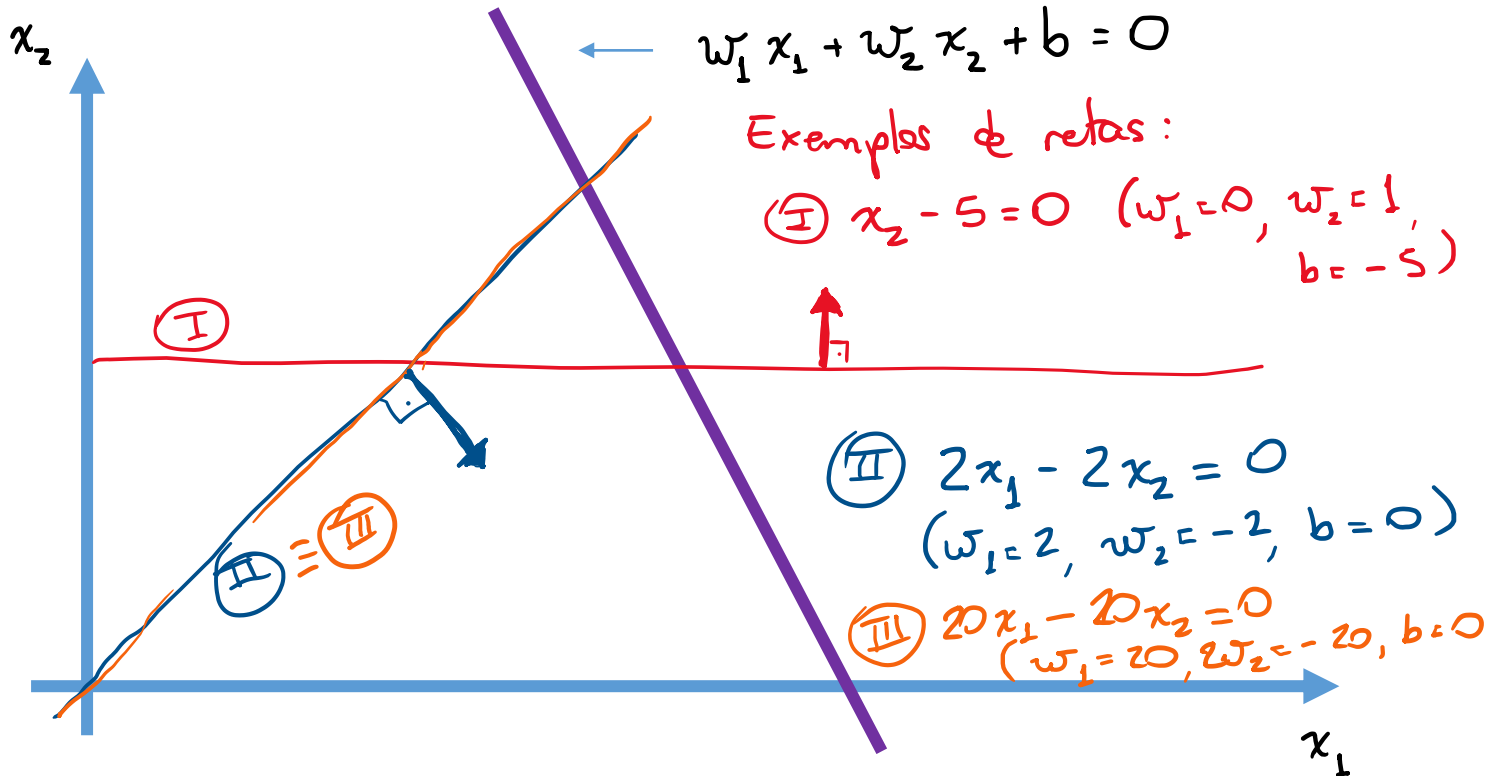
Mais problemas...



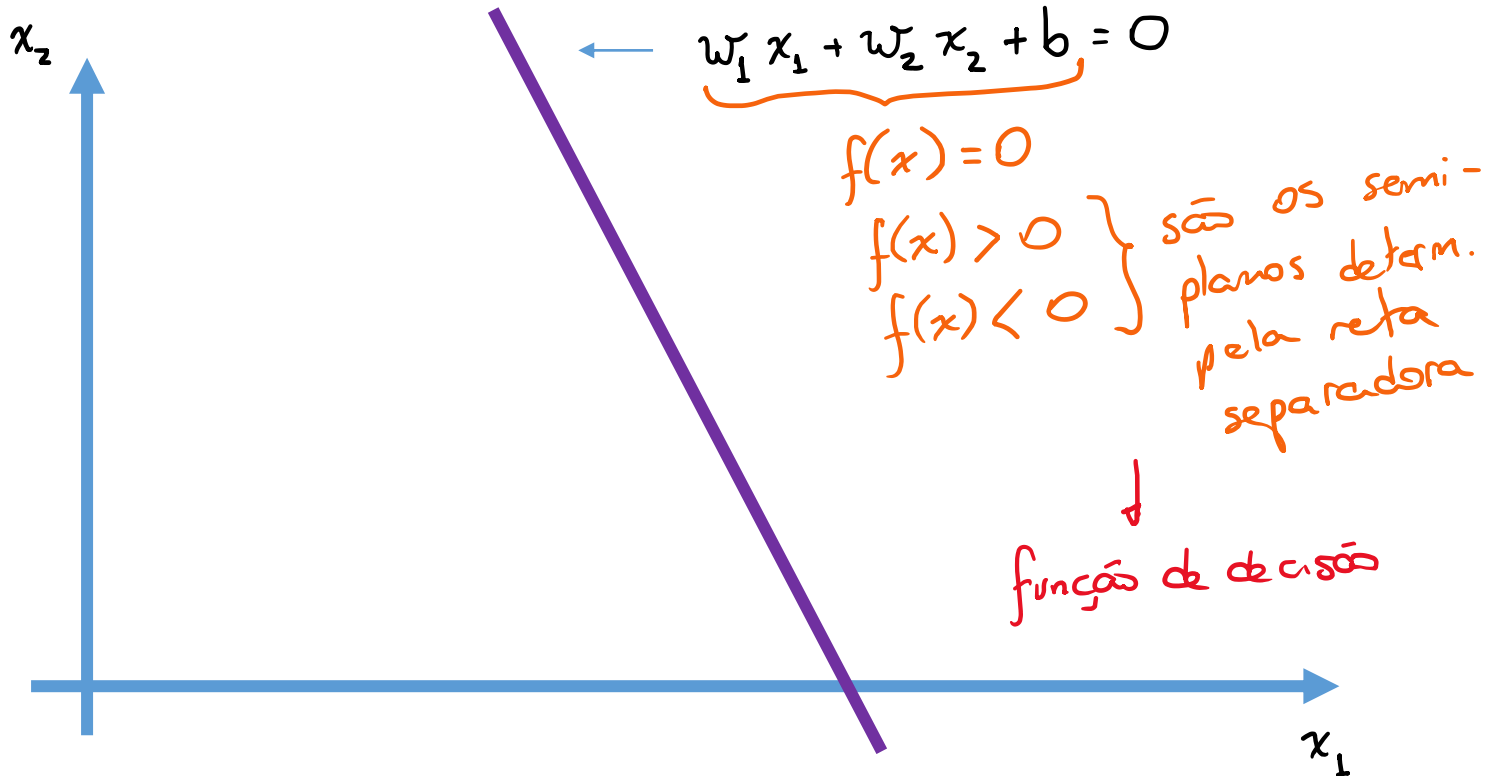
Vamos aos detalhes

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

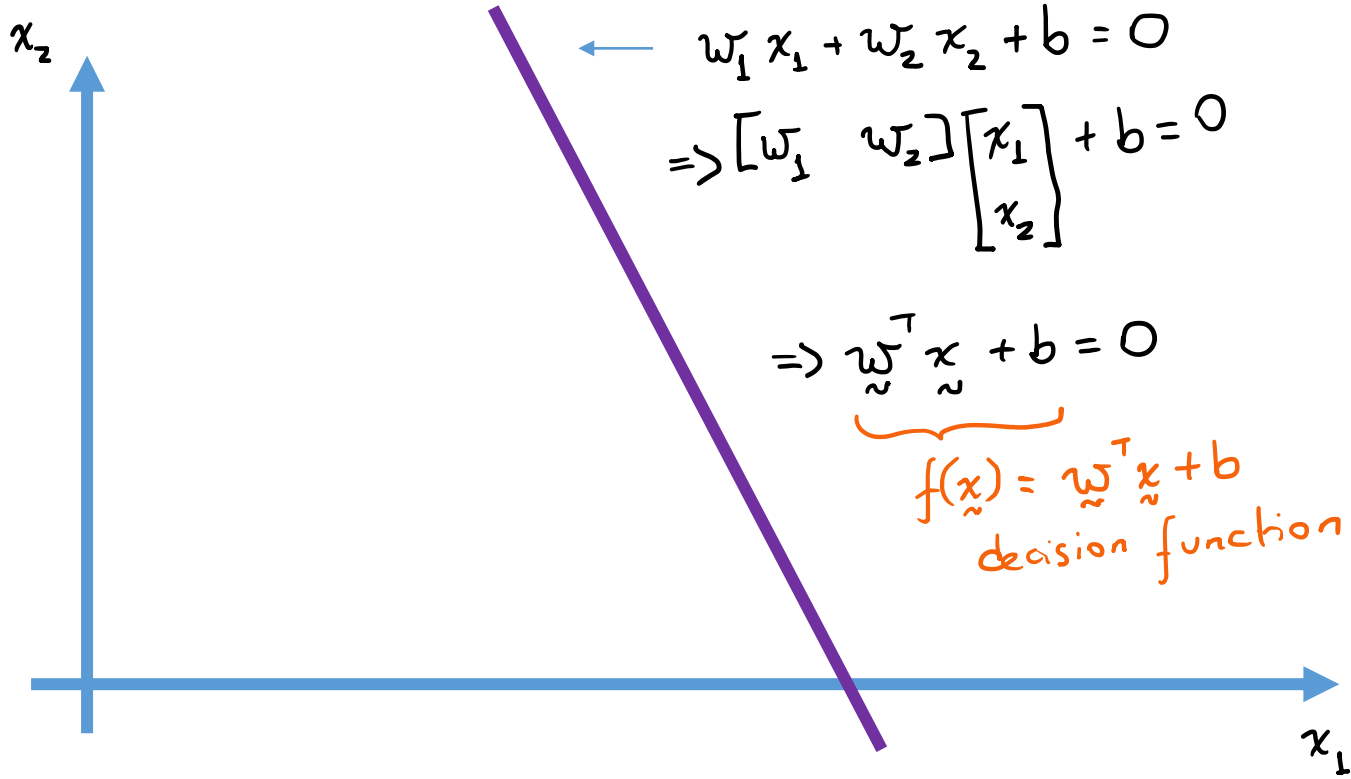
Equação da reta



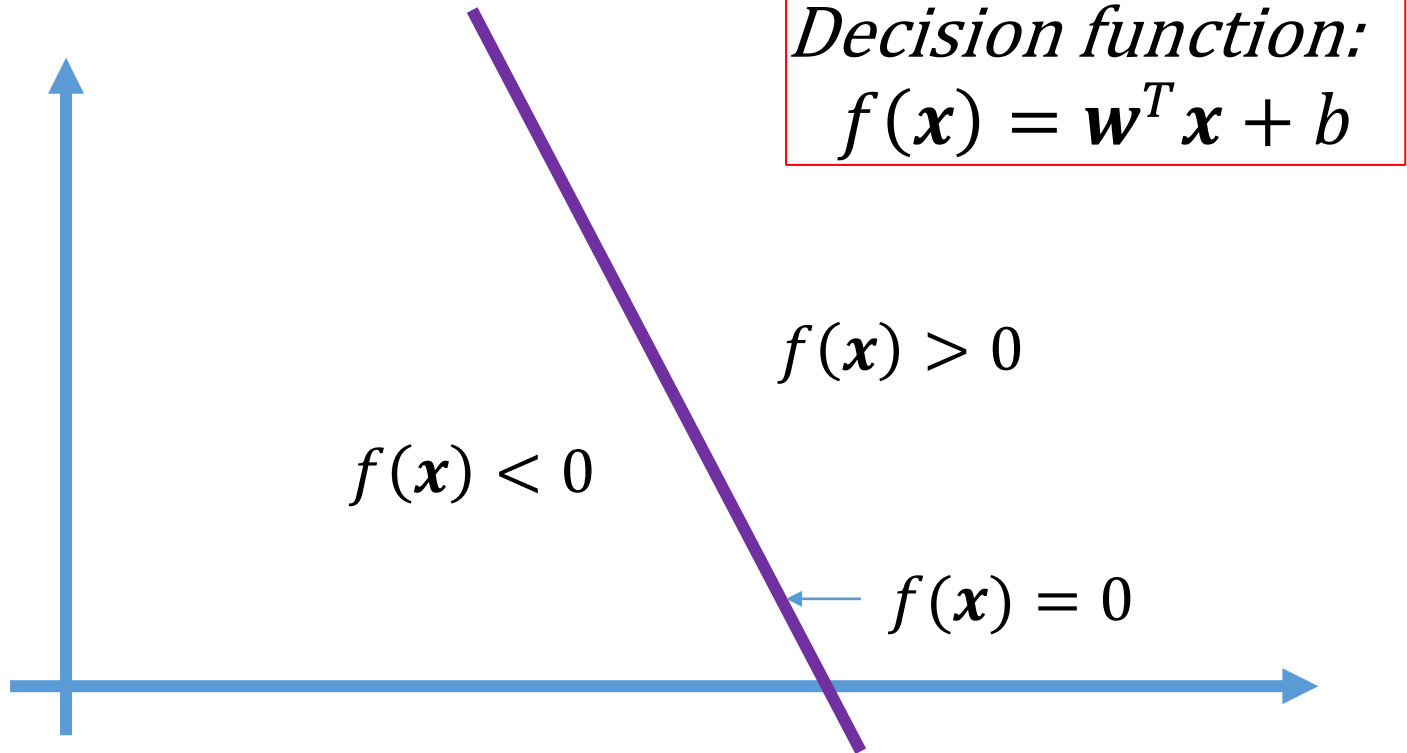
Equação da reta



Equação da reta



Equação da reta



Objetivo

Descobrir qual $f(x)$ implementa a melhor “avenida”

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

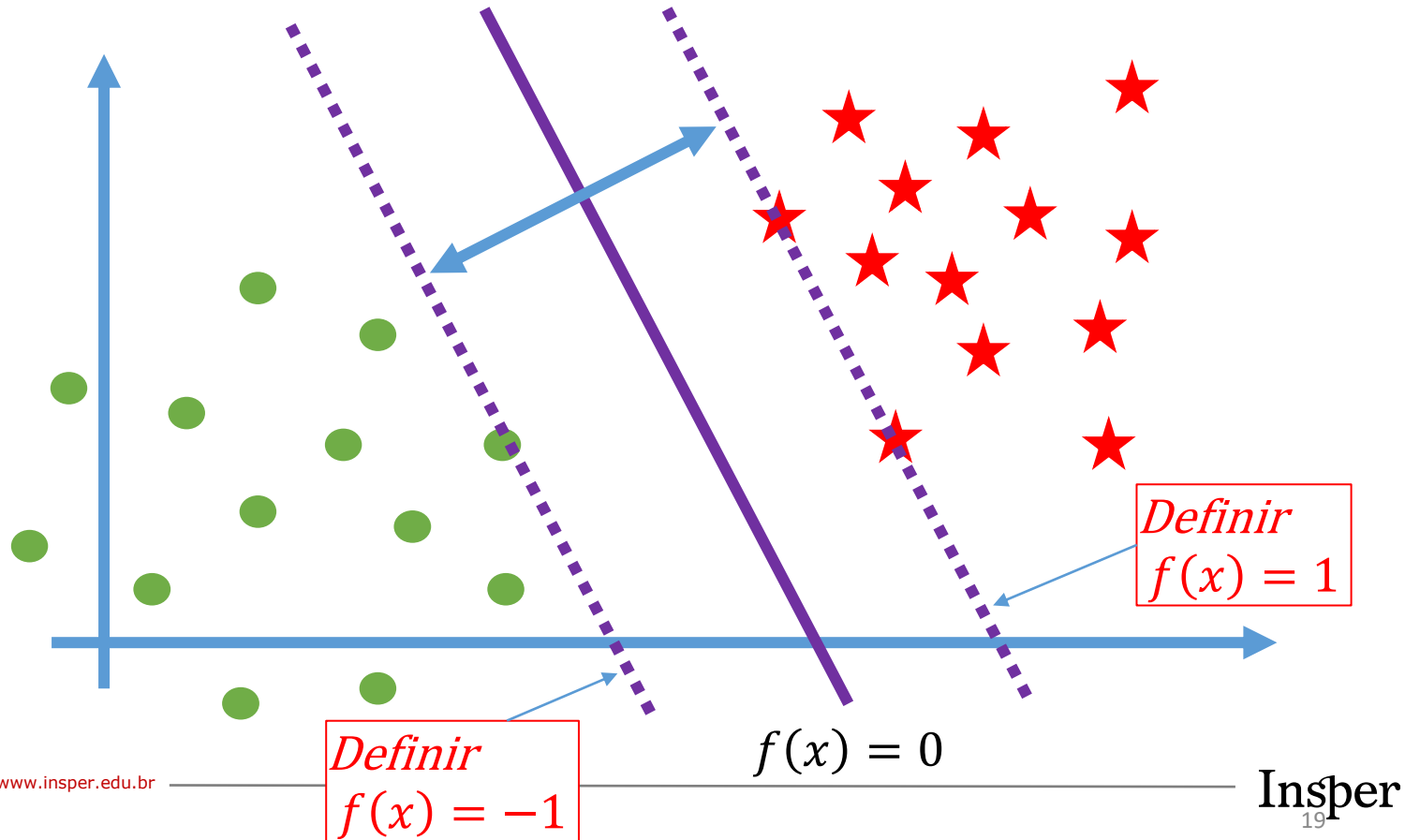
Então $g(x) = kf(x)$ também é igualmente boa!

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

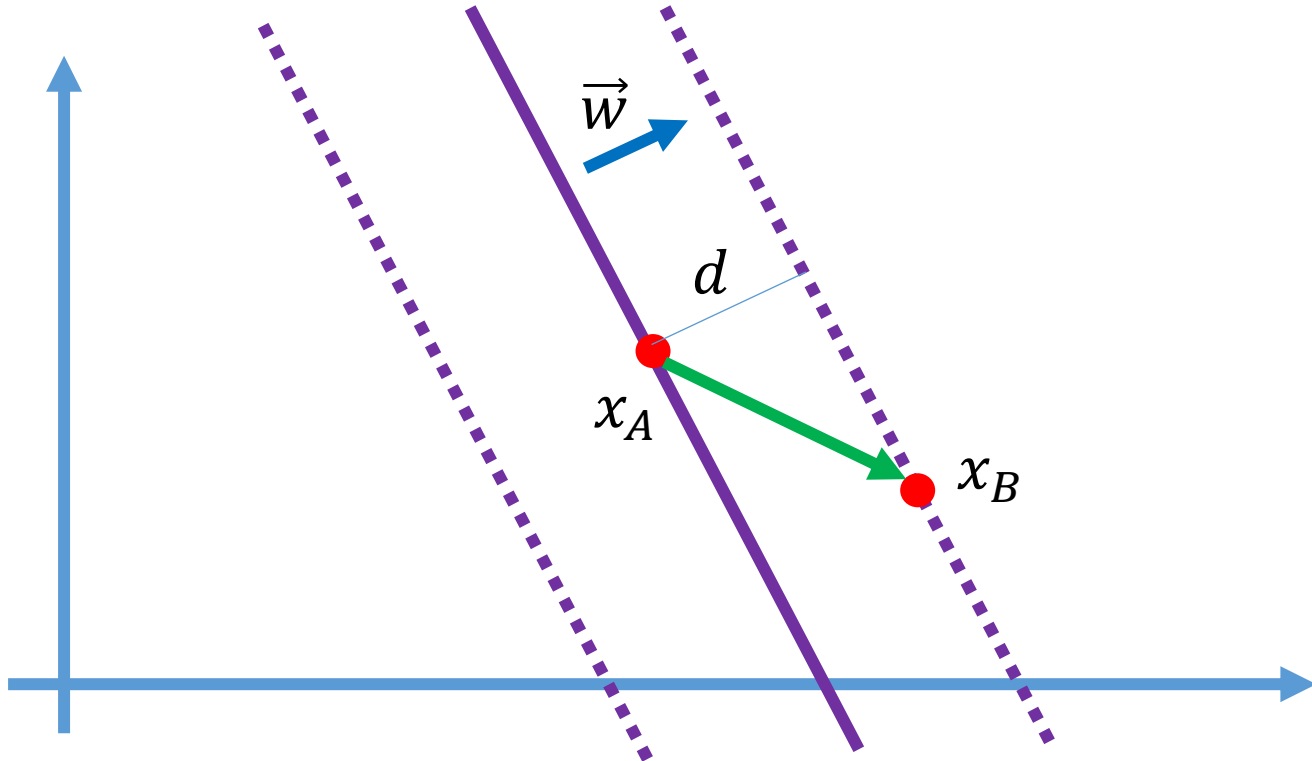
$\Rightarrow f(x)$ e $g(x)$ definem a mesma superfície de separação

AMBIGUIDADE!

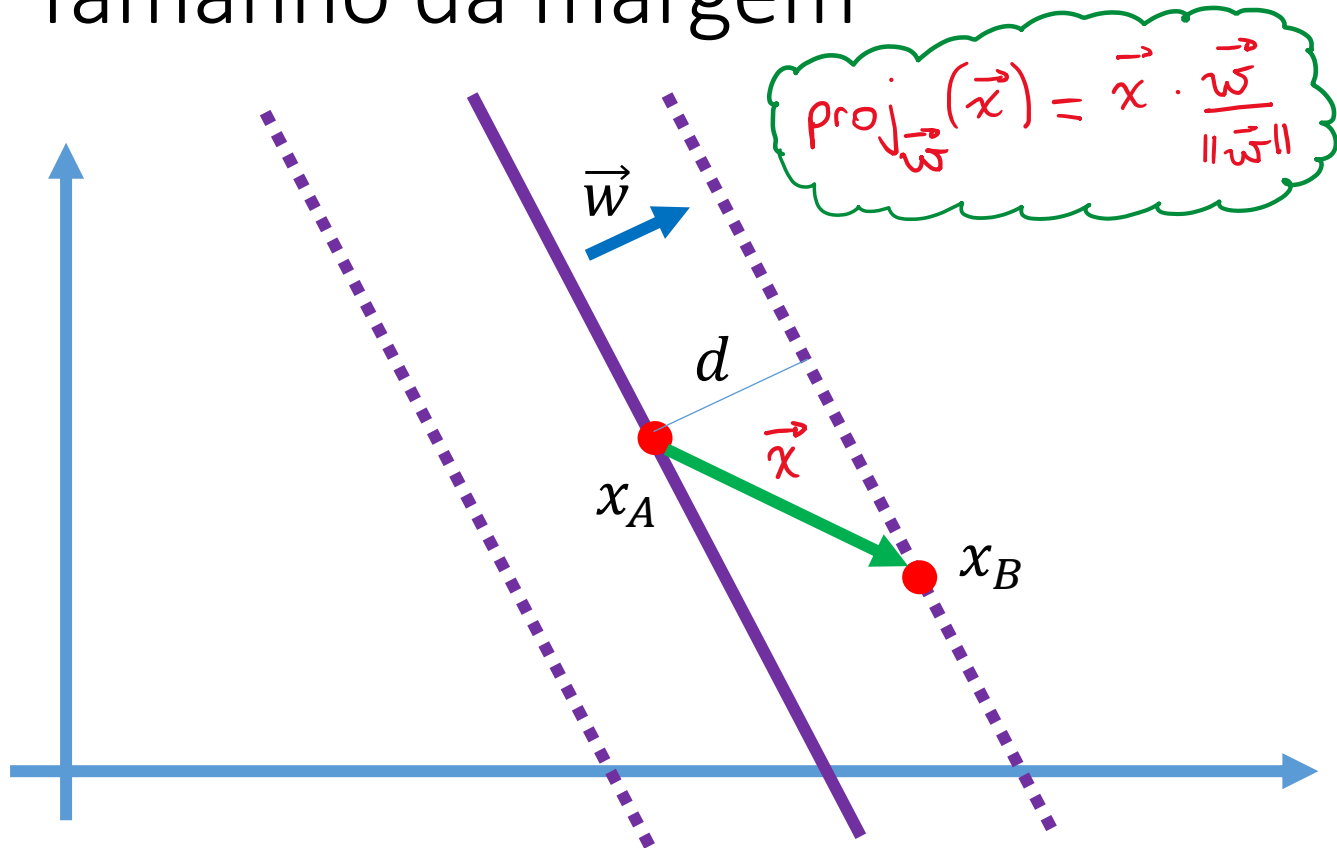
Removendo uma ambiguidade...



Tamanho da margem



Tamanho da margem



$$\left. \begin{aligned} \vec{x} &= x_B - x_A \\ d &= \vec{x} \cdot \frac{\vec{w}}{\|\vec{w}\|} \end{aligned} \right\} \Rightarrow d = \frac{\vec{w} \cdot (x_B - x_A)}{\|\vec{w}\|} = \frac{w^T (x_B - x_A)}{\sqrt{w^T w}}$$

$$= \begin{bmatrix} w_1 & \dots & w_n \end{bmatrix} \begin{bmatrix} x_B \\ \vdots \\ x_A \end{bmatrix} = w_1^2 + \dots + w_n^2$$

x_B na "calçada": $f(x_B) = 1$

$$\Rightarrow w^T x_B + b = 1 \Rightarrow w^T x_B = 1 - b$$

x_A no meio da avenida: $f(x_A) = 0$

$$\Rightarrow w^T x_A + b = 0 \Rightarrow w^T x_A = -b$$

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

$$d = \frac{1}{\sqrt{w^T w}}$$

Tamanho da margem

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

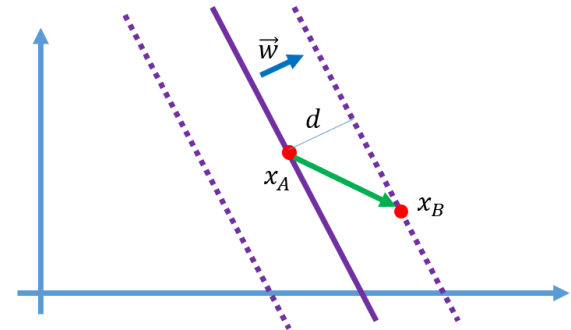
Em B: $f(x_B) = 1 \Rightarrow w^T x_B + b = 1 \Rightarrow w^T x_B = 1 - b$

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

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

Portanto:

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



Ou seja: maximizar d equivale a minimizar $w^T w$

Problema de otimização da SVM

minimizar $w^T w$

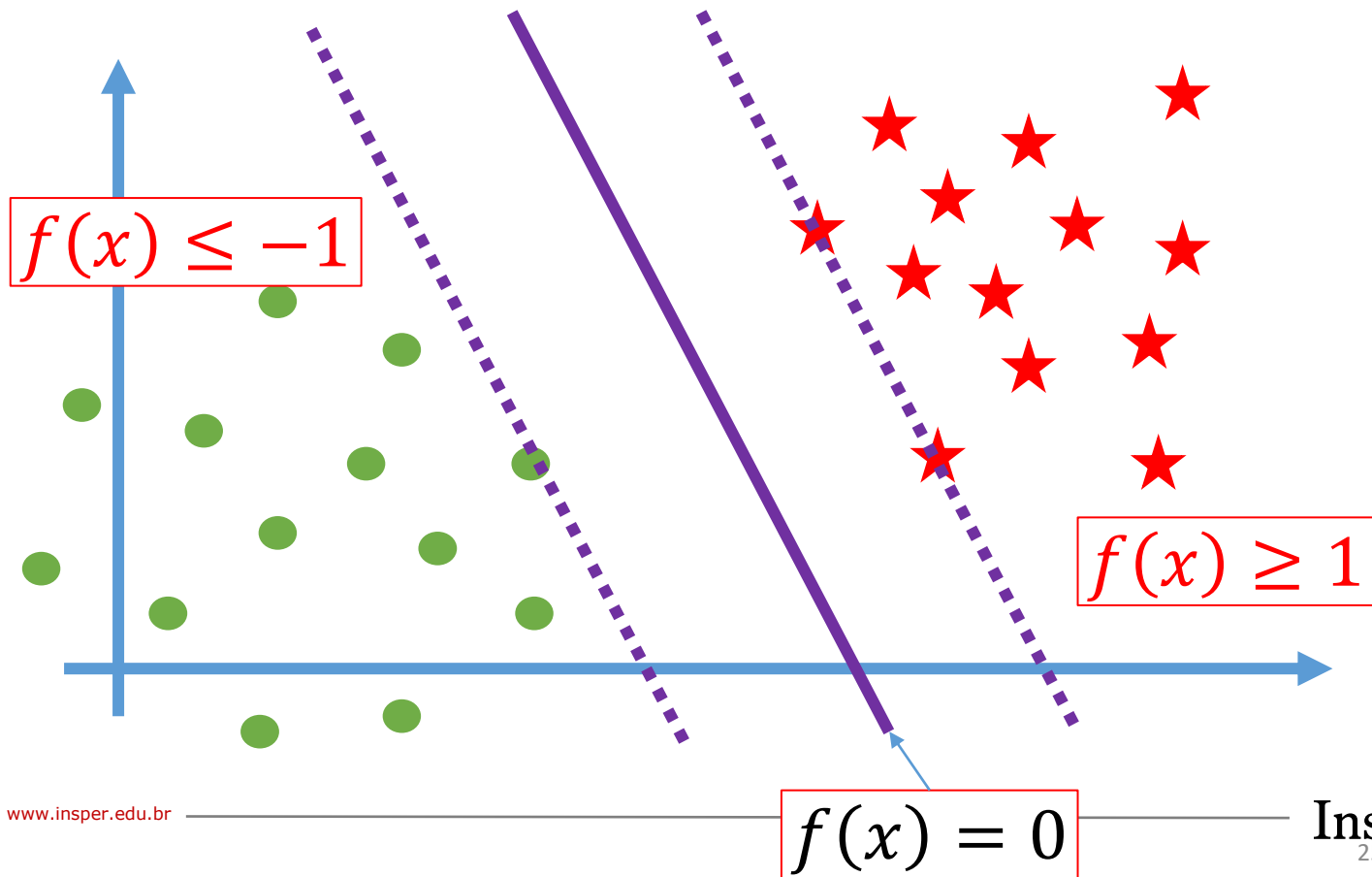
sujeito a: respeitar a "regra da calçada"

minimizar $w^T w$

sujeito a: $w^T x_i + b \geq 1$ p/as amostras "positivas"
 ≤ -1 p/as amostras "negativas"

\uparrow
i-ésima amostra de treinamento

Critério: pontos fora da “avenida”



Critério: pontos fora da “avenida”

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

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?

$$t_i f(x_i) : \quad i) \quad f(x_i) \geq 1 \Rightarrow t_i = 1 \Rightarrow t_i f(x_i) \geq 1$$

$$ii) \quad f(x_i) \leq -1 \Rightarrow t_i = -1 \Rightarrow t_i f(x_i) \geq 1$$

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 $i = 1, 2, \dots, m$

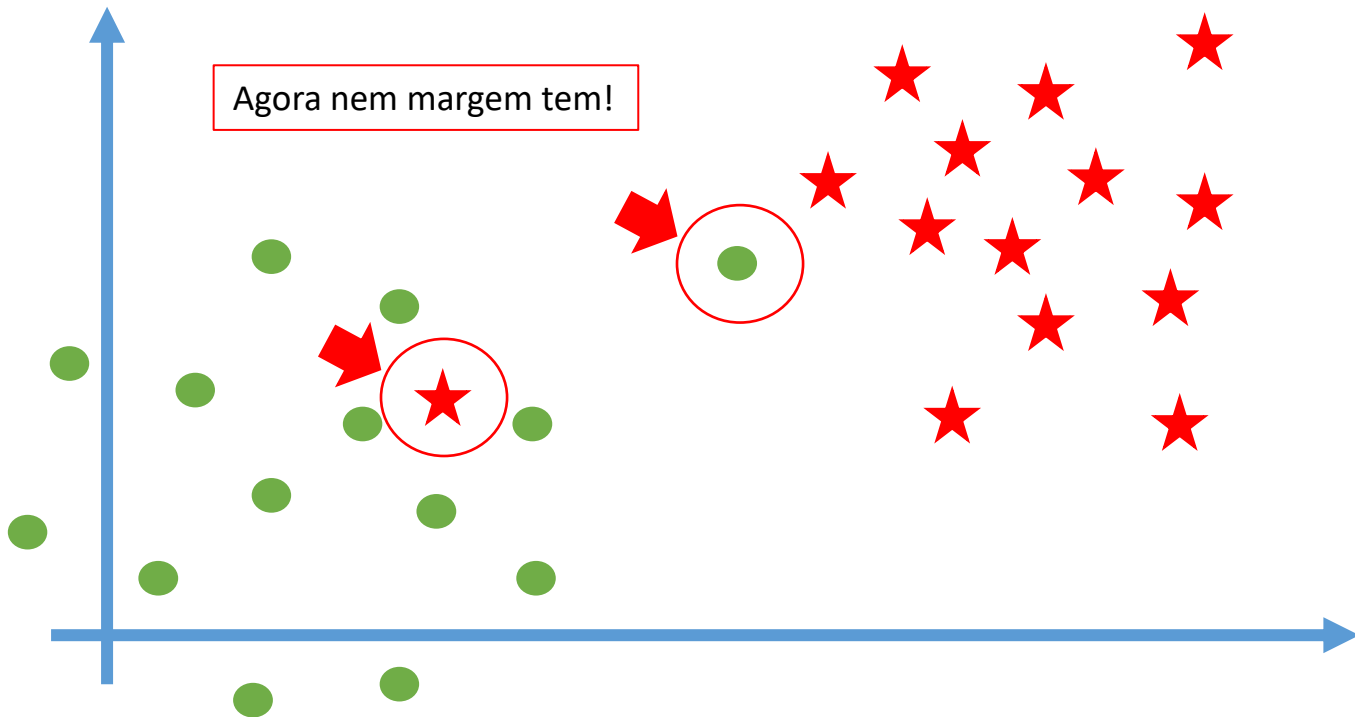
Respeitar o critério de
“pontos fora da margem”

Formulação do problema
de otimização

Problema de otimização
quadrática

TEM ALGORITMO EFICIENTE!

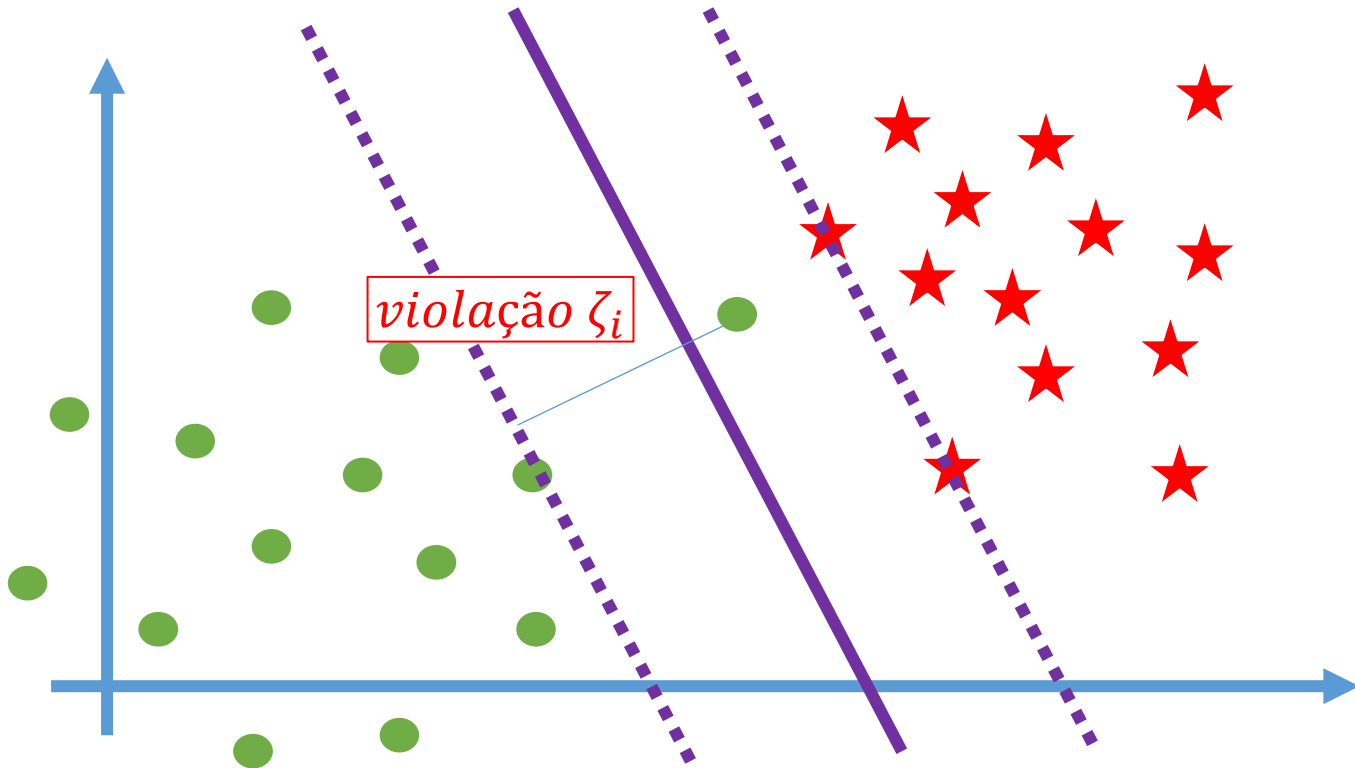
E como fica esse caso?



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

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

Hiperparâmetro! 

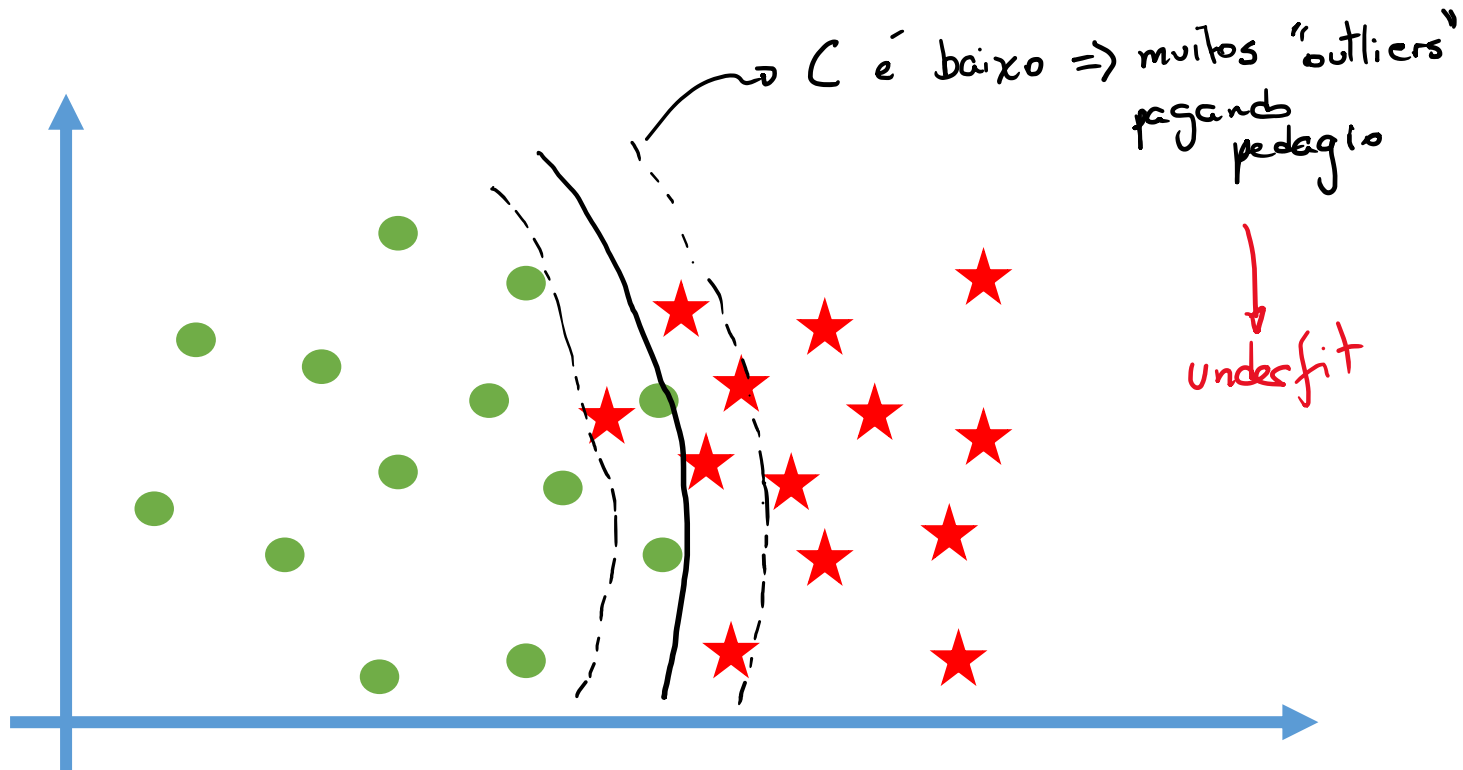
Maximizar a margem de classificação
com penalidade

$$\text{sujeito a } t_i(w^T x_i - b) \geq (1 - \zeta_i) \text{ e } \zeta_i \geq 0$$

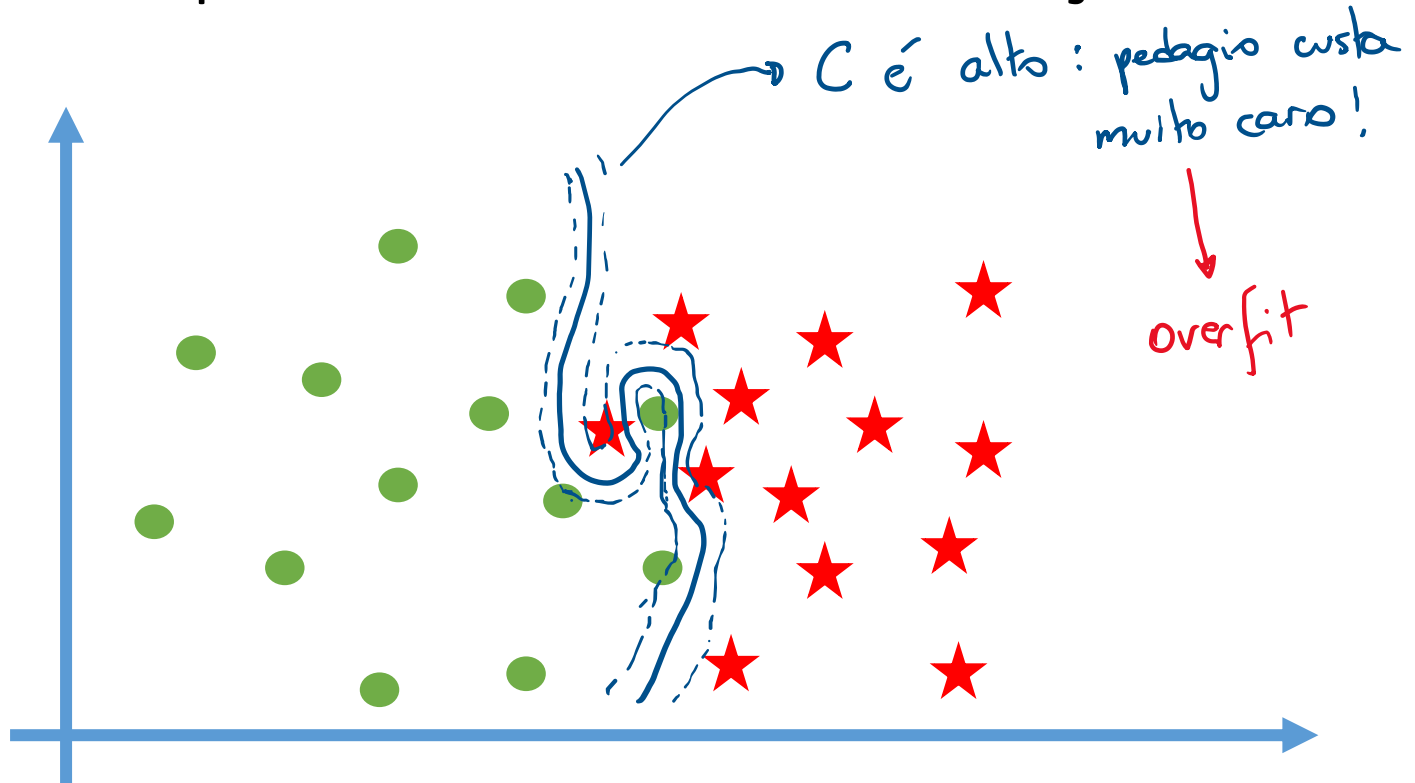
para $i = 1, 2, \dots, m$

Respeitar o critério de
“pontos fora da margem”
com permissão de outliers

Um problema de classificação



Um problema de classificação



The background of the slide is white, featuring several concentric, partial arcs in red and grey. These arcs are of varying radii and are scattered across the frame, creating a dynamic, abstract pattern. The word "Insper" is centered on the left side of the slide, enclosed within a large, thin red arc.

Insper

Pensando um pouco mais sobre otimização

Problema original:

$$\text{minimizar } \frac{1}{2} w^T w \quad \text{sujeito a } t_i(w^T x_i - b) \geq 1, \\ \text{para } i = 1, 2, \dots, m$$

Multiplicadores de Lagrange funciona aqui?
(afinal tem desigualdade...)

Multiplicadores de Lagrange com desigualdade

Problema original:

$$\text{minimizar } \frac{1}{2} w^T w \quad \text{sujeito a } t_i(w^T x_i - b) \geq 1, \\ \text{para } i = 1, 2, \dots, m$$



$$\text{minimizar } L = \frac{1}{2} w^T w - \sum_{i=1}^m \alpha_i (t_i(w^T x_i - b) - 1) \\ \text{sujeito a } \alpha_i \geq 0 \text{ para } i = 1, 2, \dots, m$$

Condições Karush-Kuhn-Tucker

Um mínimo do Lagrangiano que respeita as condições:

- $t_i(w^T x_i - b) \geq 1$
- $\alpha_i \geq 0$
- $\alpha_i = 0$ se $t_i(w^T x_i - b) > 1$, caso contrario $t_i(w^T x_i - b) = 1$

é um ponto ótimo do problema original

Eliminando w e b : forma dual

Tirando a derivada de L em relação a w e a b e igualando a zero temos duas equações:

$$w = \sum_{i=1}^m \alpha_i t_i x_i$$

e

$$\sum_{i=1}^m \alpha_i t_i = 0$$

Eliminando w e b : forma dual

Substituindo as expressões anteriores em L temos o **problema de otimização dual**:

$$\text{minimizar } \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j t_i t_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^m \alpha_i$$

sujeito a $\alpha_i \geq 0$ para $i = 1, 2, \dots, m$

Uma vez encontrados os α_i , substituir nas equações anteriores para achar w e b

O famoso “kernel trick”

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j t_i t_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^m \alpha_i$$



$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j t_i t_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) - \sum_{i=1}^m \alpha_i$$

O famoso “kernel trick”

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j t_i t_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) - \sum_{i=1}^m \alpha_i$$



$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j t_i t_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^m \alpha_i$$

Teorema de Mercer

$K(a,b)$ é contínua e simétrica

então existe $\phi(\cdot)$ tal que $K(a,b) = \phi(a)^T \phi(b)$

\Rightarrow SVM com kernel não-linear

|||

SVM linear em um espaço aumentado
que eu nem sei qual é!

Equation 5-10. Common kernels


Linear: $K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}$

Polynomial: $K(\mathbf{a}, \mathbf{b}) = (\gamma \mathbf{a}^\top \mathbf{b} + r)^d$

Gaussian RBF: $K(\mathbf{a}, \mathbf{b}) = \exp(-\gamma \|\mathbf{a} - \mathbf{b}\|^2)$

Sigmoid: $K(\mathbf{a}, \mathbf{b}) = \tanh(\gamma \mathbf{a}^\top \mathbf{b} + r)$

Class	Time complexity	Out-of-core support	Scaling required	Kernel trick
LinearSVC	$O(m \times n)$	No	Yes	No
SGDClassifier	$O(m \times n)$	Yes	Yes	No
SVC	$O(m^2 \times n)$ to $O(m^3 \times n)$	No	Yes	Yes

USE  StandardScaler

Regressão

