Aprendizado de Máquina:
Support Vector Machines (SVM)

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Introdução

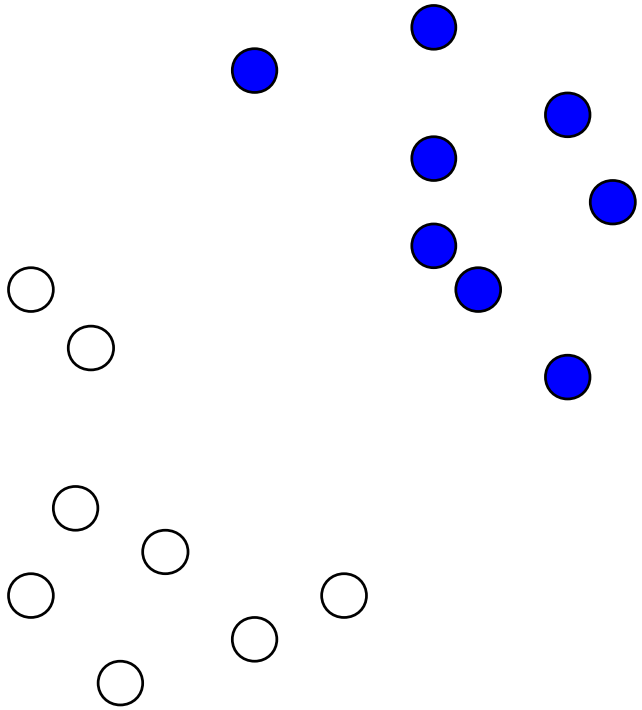
- Método supervisionado de aprendizado de máquina
- Classificação em dois grupos
  - Classificação de múltiplas classes não é uma limitação, pois pode-se construir uma SVM para cada classe
- Apresenta resultados melhores que muitos métodos populares de classificação

Introdução (2)

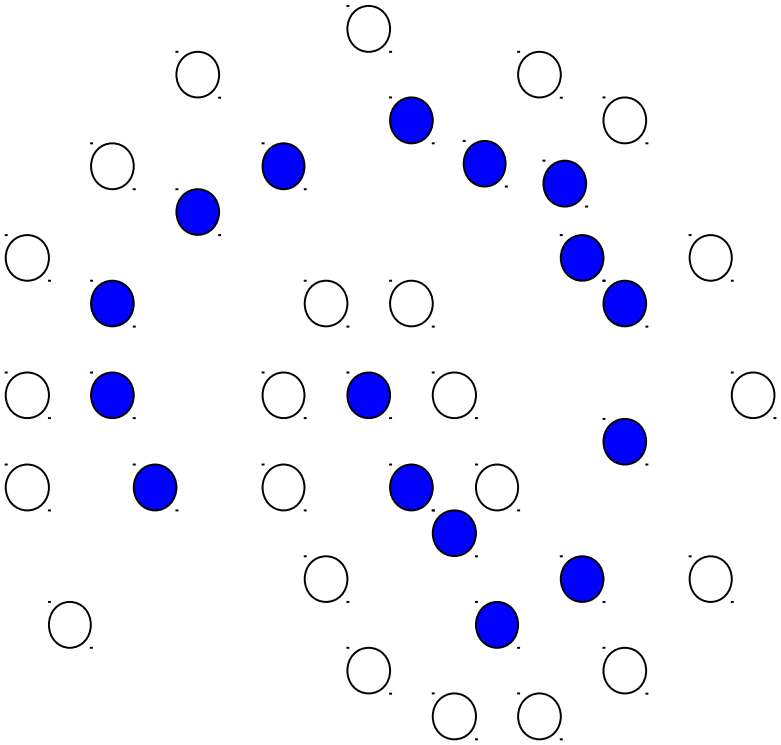
- [Vapnik et al, 1992] Primeiro artigo
- [Vapnik et al, 1998] Definição detalhada
- 1968: base matemática (teoria de Lagrange)

Motivação da SVM

- Como separar as duas classes?

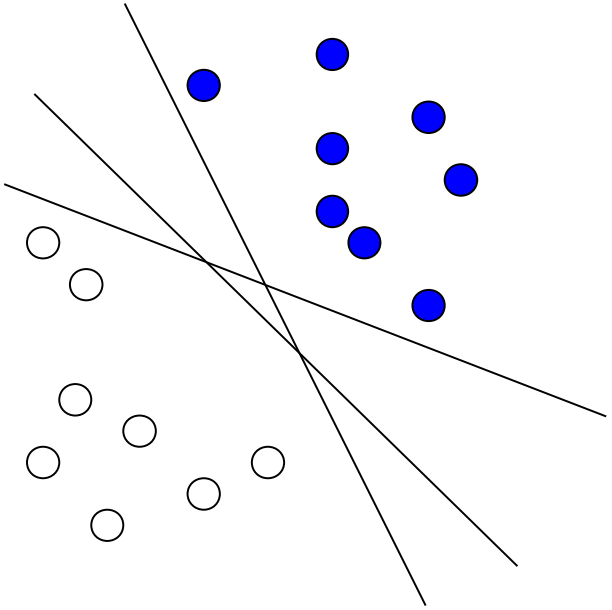


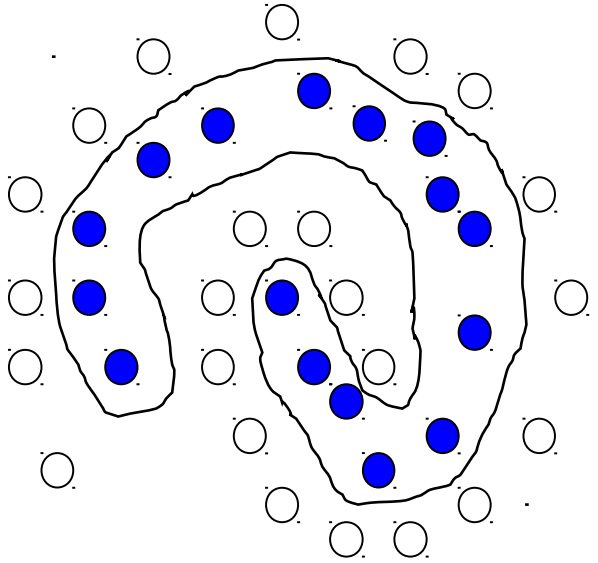
- Como separar as duas classes?



Motivação da SVM (2)

- Reta / Plano / Hiperplano?

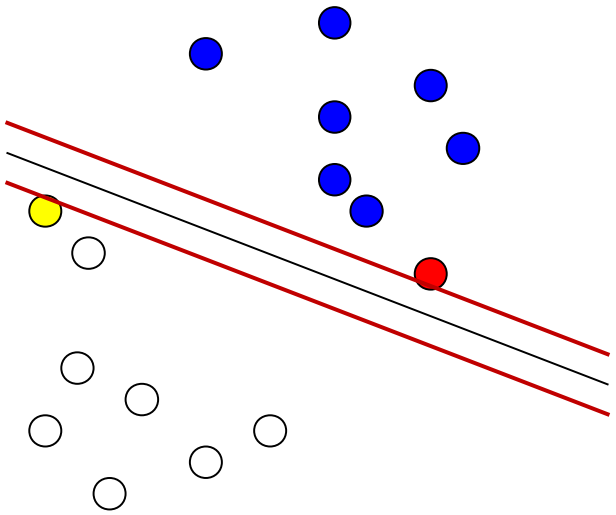




- Qual o hiperplano ótimo?
  Menor erro de classificação

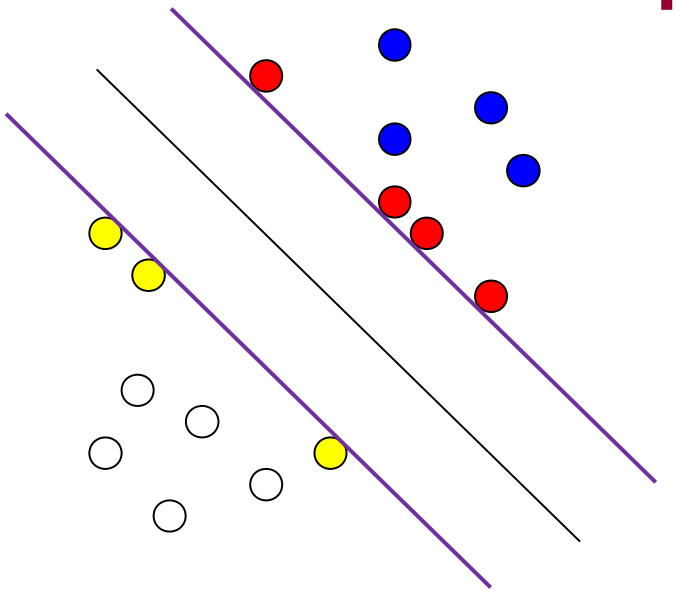
Conceitos de SVM

- Qual o **hiperplano** ótimo?
  - Menor erro de classificação
  - Maior **margem**
    - Distância entre **vetores de suporte** e o hiperplano

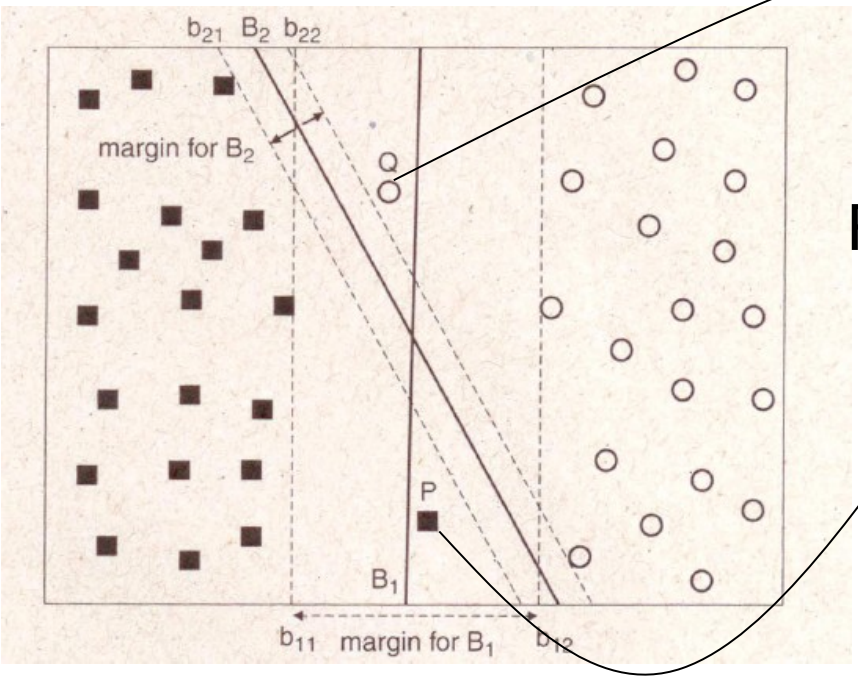


Conceitos de SVM (2)

- Qual o **hiperplano** ótimo?
  - Menor erro de classificação
  - Maior **margem**
    - Distância entre **vetores de suporte** e o hiperplano



Casos a tratar (2)



Ruídos ?

- B1 ainda é o melhor separador !
- Mas B2 é o que seria produzido usando uma técnica própria para dados "separáveis"
- **Enfoque Soft Margin:** Produz fronteira que tolera algumas "exceções" na separação.

Como Funciona para Dados Linearmente Separáveis?

- Dados de treinamento
  - Tuplas no formato (x_1, x_2, ..., x_n, y)
    - Atributos x_i
    - Classe y (+1, -1)
- Conjunto dito linearmente separável, se existir um hiperplano H (no espaço de entrada) que separe as tuplas de classes diferentes
- Determinar os **vetores de suporte**
- Encontrar o **hiperplano** ótimo
  - Com maior **margem**

O Hiperplano (H)

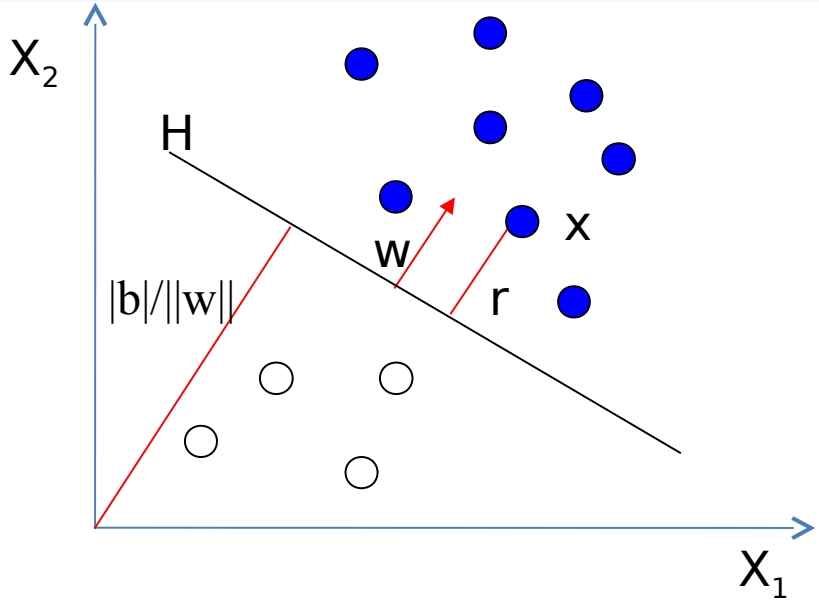
- Pontos que pertencem a h satisfazem a equação

$$\hat{w} \cdot \hat{x} + b = 0$$

- $\hat{w}$: vetor normal a h

$\hat{w} = w_1, w_2, ..., w_n$

- **Convenção**: pontos no hiperplano separador pertencem a classe positiva



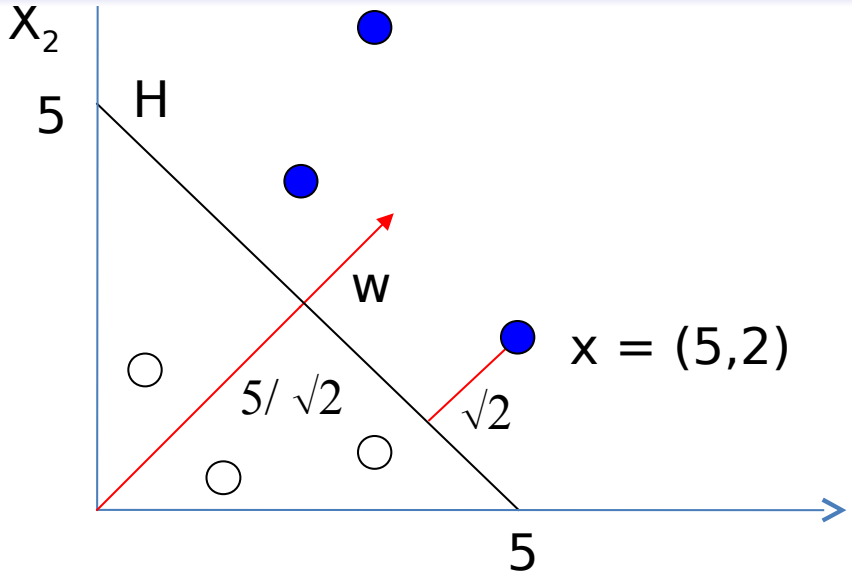
O Hiperplano (H) (2)

- A distância r entre um ponto x e o hiperplano H é
  - $d = (\hat{w} \cdot \hat{x} + b) / ||\hat{w}||$
- $||w||$ é a norma euclidiana de w
  - $\sqrt{} (\hat{w} \cdot \hat{w}) = \sqrt{} (\hat{w}_1^2 + ... + \hat{w}_n^2)$
- $|b|/||w||$ é a distância perpendicular de H até a origem

O Hiperplano (H) (3)

- Orientação de $\hat{w}$
  - Define o lado do plano em que os pontos pertencem a classe +1
- $b > 0$ (origem no lado positivo)
- $b < 0$ (origem no lado negativo)
- $b = 0$ (origem pertence ao plano)

Hiperplano



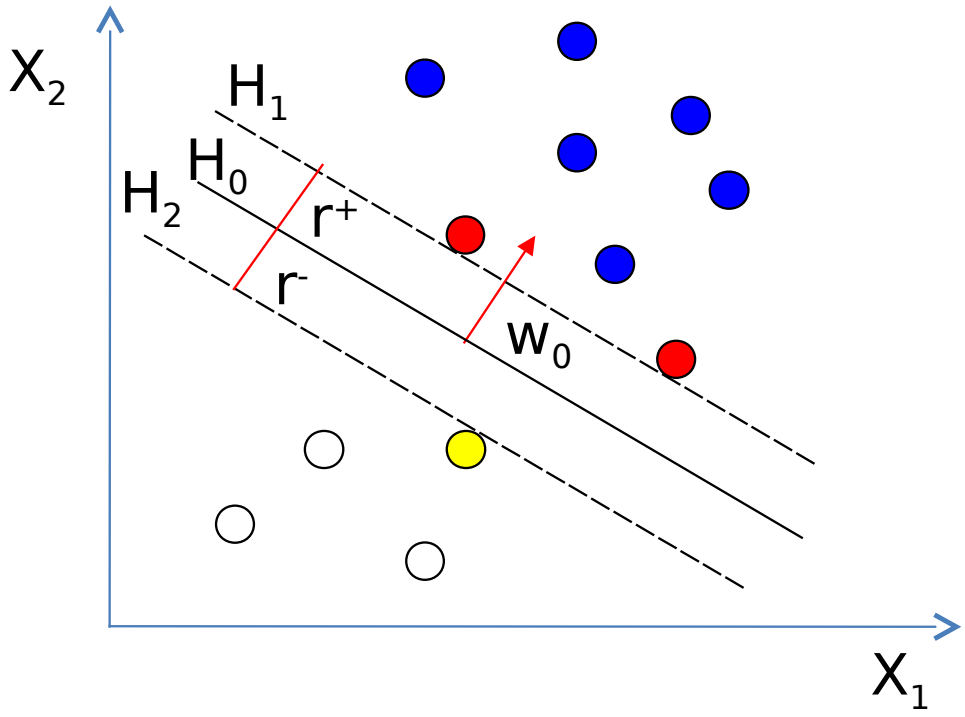
- H: $\hat{w} \cdot \hat{x} + b = 0$
  H: $w_1x_1 + w_2x_2 + b = 0$
- Aplicando os pontos (5,0) e (0,5)
  $5w_1 + b = 0$
  $5w_2 + b = 0$
- Isolando b
  $5w_1 = 5w_2$ ($w_1 = w_2$)
- Escolhendo arbitrariamente $w_1 = 1$; $b = -5$

Hiperplano (2)

- Norma de $\hat{w}$
  - $||\hat{w}|| = \sqrt{(w_1^2+w_2^2)} = \sqrt{2}$
- Distância da origem
  - $|b| / ||\hat{w}|| = 5/\sqrt{2}$
- Distância de um ponto x = (5,2) até H
  - $r = (\hat{w} . \hat{x} + b) / ||\hat{w}||$
    $r = (5w_1 + 2w_2 -5) / \sqrt{2}$
    $r = (5+2-5) / \sqrt{2}$
    $r = \sqrt{2}$

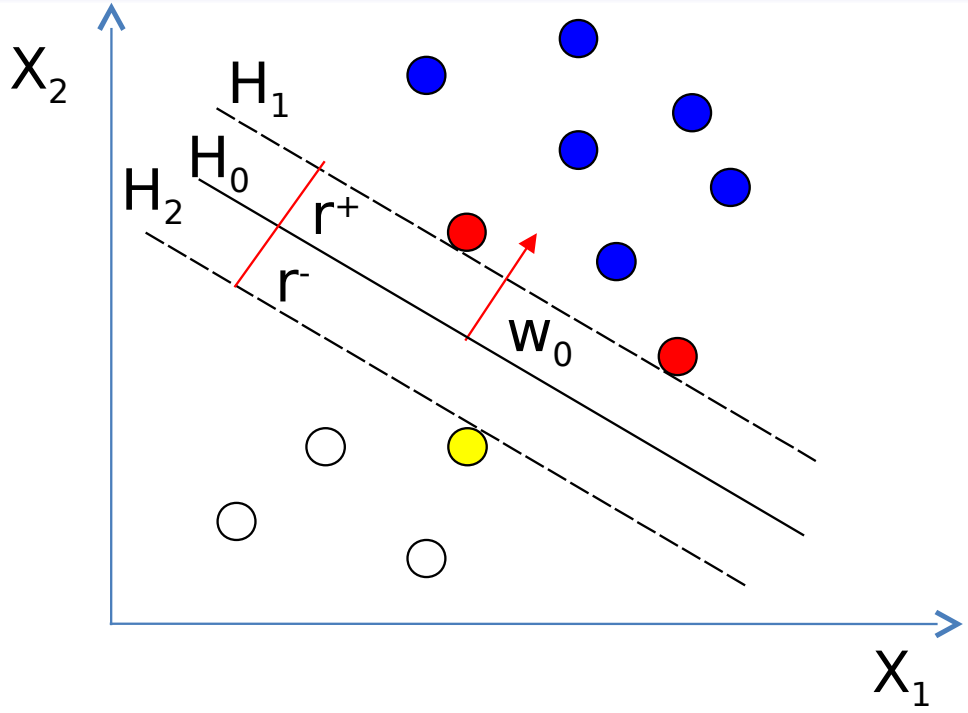
Hiperplano Ótimo

- Objetivo da SVM é encontrar ŵ e b para a maior margem



Hiperplano Ótimo (2)

- h_0: $\hat{w} \cdot \hat{x} + b = 0$
- h_1: $\hat{w} \cdot \hat{x} + b = 1$
- h_2: $\hat{w} \cdot \hat{x} + b = -1$



- Obs: existem infinitas formas de representar os três hiperplanos
- Estamos escolhendo uma conveniente

Hiperplano Ótimo (3)

- Hiperplano ótimo, $r^+ = r^-$
  - $r = 1 / ||\hat{w}||$
  - Margem $= 2 / ||\hat{w}||$
  - É aquele que possui maior margem
  - É aquele que possui menor $||\hat{w}||$

Hiperplano Ótimo (4)

- Formas de determinar o hiperplano
  - Sistema de equações (**nosso foco**)
  - Problema de otimização restrita
    - Usado em aplicações reais
    - Fora do escopo da disciplina
    - Breve explicação a seguir

Avançado: treinamento na prática

- Otimização dual ou primal :

$$\underset{\alpha}{\operatorname{argmax}} \sum_j \alpha_j - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_k y_j y_k (\mathbf{x}_j \cdot \mathbf{x}_k)$$

  - α_i : coeficientes a encontrar
  - x_i: vetores de suporte
  - y_i: classes desejadas dos vetores de suporte
  - Importante: usa **produto escalar**

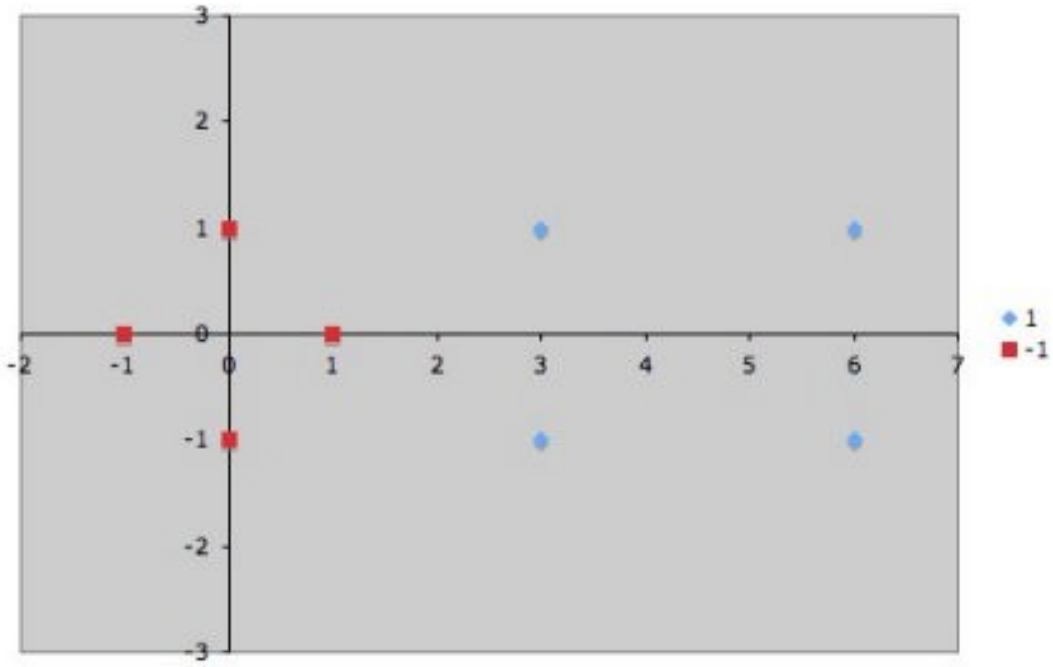
Avançado: classificação na prática

$$h(\mathbf{x}) = \text{sign}\left(\sum_j \alpha_j y_j (\mathbf{x} \cdot \mathbf{x}_j) - b\right)$$

- α_i : coeficientes encontrados no passo anterior
- x: instância a ser classificada
- x_j: vetores de suporte
- y_i: classes desejadas dos vetores de suporte
- b: primeira somatória da equação anterior
- *sign*: sinal (+ ou -)
- Importante: usa **produto escalar**

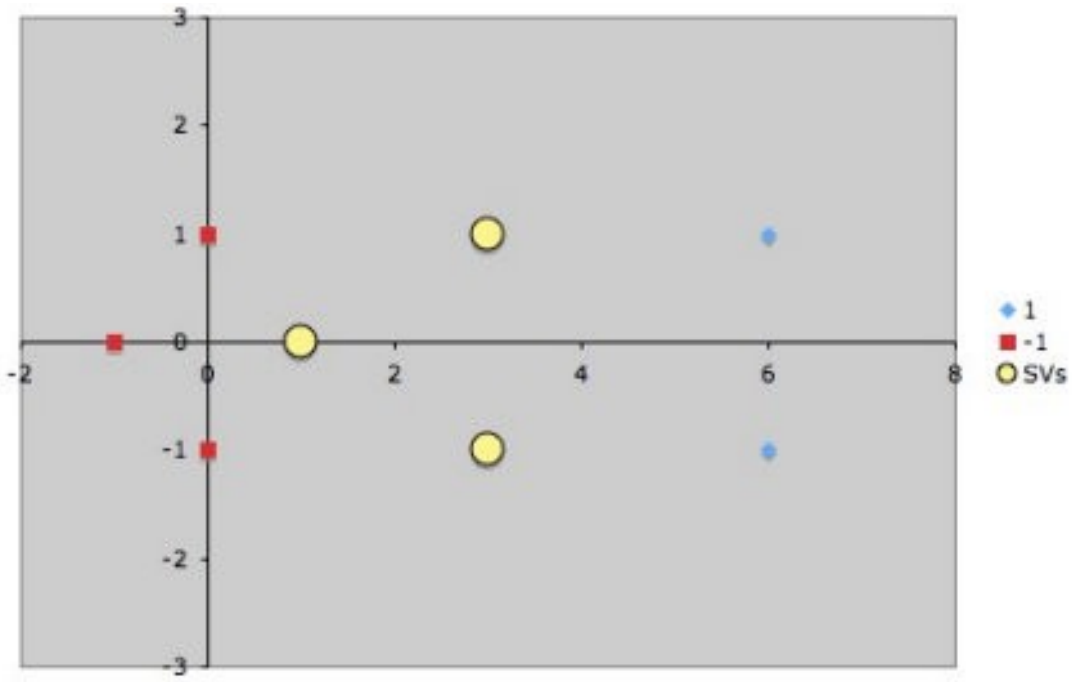
Exemplo

- -1, 0, -1
- 0, -1, -1
- 0, 1, -1
- 1, 0, -1
- 3, -1, +1
- 3, 1, +1
- 6, -1, +1
- 6, 1, +1



Exemplo (2)

- -1, 0, -1
- 0, -1, -1
- 0, 1, -1
- **1, 0, -1**
- **3, -1, +1**
- **3, 1, +1**
- 6, -1, +1
- 6, 1, +1



Exemplo (3)

- H_1: $w \cdot x + b = 1$
  H_2: $w \cdot x + b = -1$

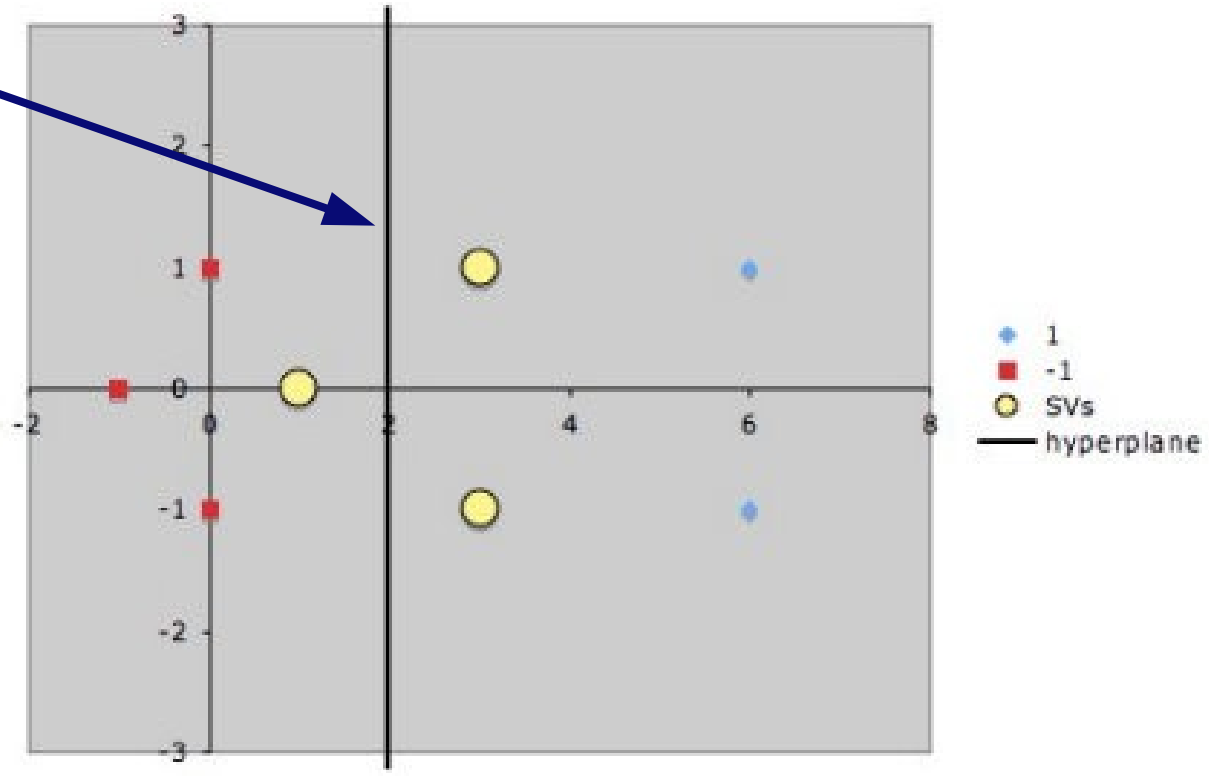
(1, 0) → -1
(3, -1) → +1
(3, 1) → +1

- $1w_1 + 0w_2 + b = -1$
  $\rightarrow b = -1 - w_1$
- $3w_1 - 1w_2 + b = 1$
  $\rightarrow w_2 = 3w_1 - 1 - w_1 - 1$
  $\rightarrow w_2 = 2w_1 - 2$
- $3w_1 + 1w_2 + b = 1$
  $\rightarrow 3w_1 + 2w_1 - 2 - 1 - w_1 = 1$
  $\rightarrow w_1 = 1$
  $\rightarrow b = -2$
  $\rightarrow w_2 = 0$

Exemplo (4)

- $(1, 0) \cdot (x_1, x_2) - 2 = 0$

  $x_1 = 2$



Exemplo (5)

- H: (1, 0) · x - 2 = 0
  H: $x_1 + 0x_2 - 2 = 0$

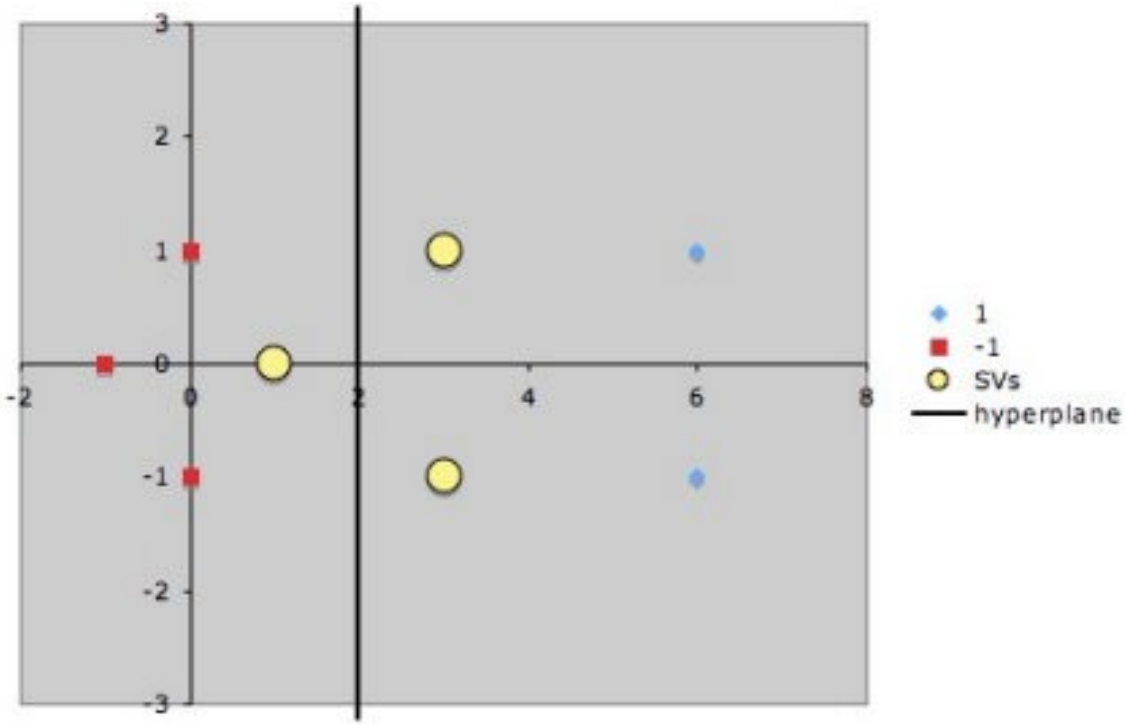
- Dados de Teste

  (4, 2), (1.5, 0.5), (0, -2)

  4 - 2 = 2 [+1]

  1.5 - 2 = -0.5 [-1]

  0 - 2 = -2 [-1]



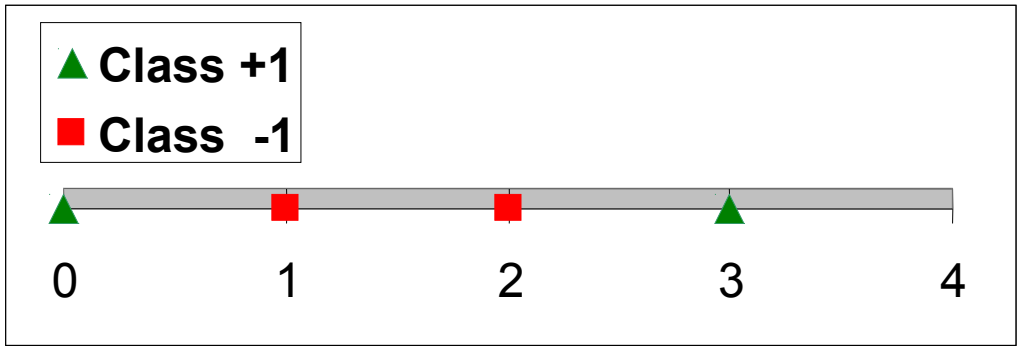
Como funciona para dados linearmente inseparáveis?

- Mapeamento do espaço de características para uma dimensão maior
- Vetores de entrada são mapeados de forma não linear
- Após transformado, o novo espaço de características deve ser passível de separação linear

Exemplo

- Como separar as duas classes com apenas um ponto?

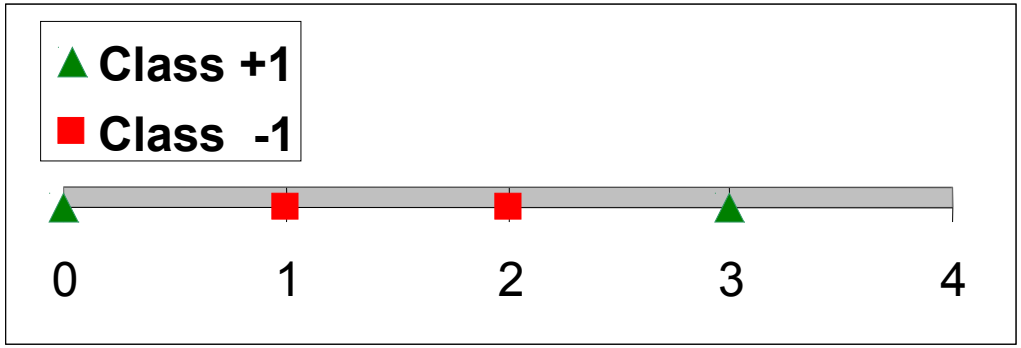
X_1	Class
0	+1
1	-1
2	-1
3	+1



Exemplo (2)

- SVM usa uma função não linear sobre os atributos do espaço de características inicial

X_1	Class
0	+1
1	-1
2	-1
3	+1

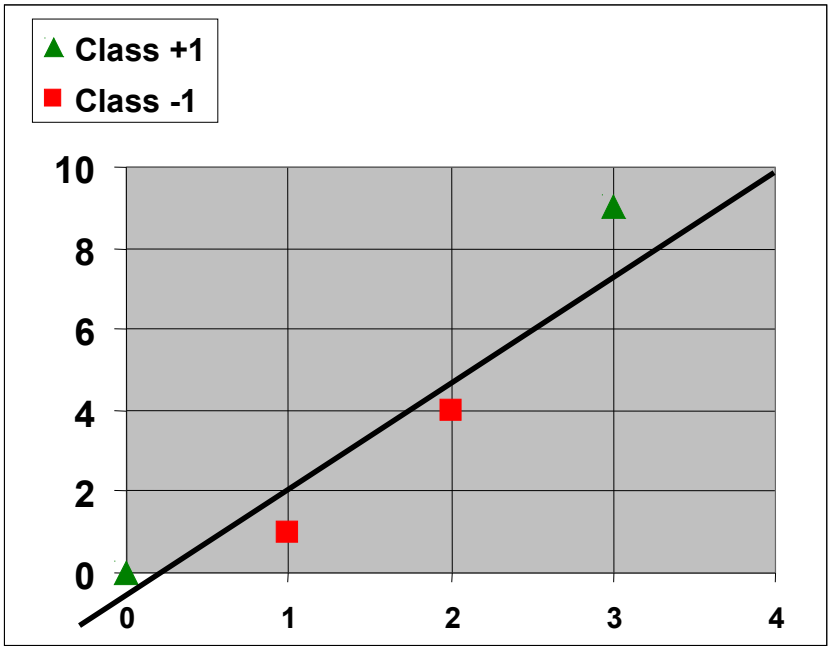


- $\Phi(x_1) = (x_1, x_1^2)$
- Esta função torna o problema bidimensional

Exemplo (3)

- SVM usa uma função não linear sobre os atributos do espaço de características inicial

X_1	X_1^2	Class
0	0	+1
1	1	-1
2	4	-1
3	9	+1



- $\Phi(x_1) = (x_1, x_1^2)$
- Esta função torna o problema bidimensional e os dados linearmente separáveis

Exemplo (4)

- $w \cdot x + b = +1$

  $w_1x_1 + w_2x_2 + b = +1$

  $0w_1 + 0w_2 + b = +1 \quad \rightarrow b = 1$

  $3w_1 + 9w_2 + b = +1$

X_1	X_1^2	Class
0	0	+1
1	1	-1
2	4	-1
3	9	+1

Exemplo (5)

X_1	X_1^2	Class
0	0	+1
1	1	-1
2	4	-1
3	9	+1

- $w \cdot x + b = -1$

  $w_1x_1 + w_2x_2 + b = -1$

  $1w_1 + 1w_2 + b = -1$

  $2w_1 + 4w_2 + b = -1$

substituindo b e após w_1

$\rightarrow w_1 = -2 - w_2$

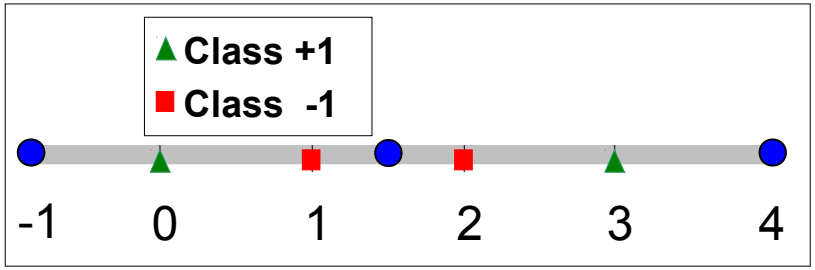
$\rightarrow -4 - 2w_2 + 4w_2 + 1 = -1$

- $w \cdot x + b = 0$ $\quad w_2 = 1$ e $w_1 = -3$

  $w_1x_1 + w_2x_2 + b = 0 \rightarrow -3x_1 + x_2 + 1 = 0$

Exemplo (6)

- **H: $-3x_1 + x_2 + 1 = 0$**
- Dados de Teste (1.5), (-1), (4)



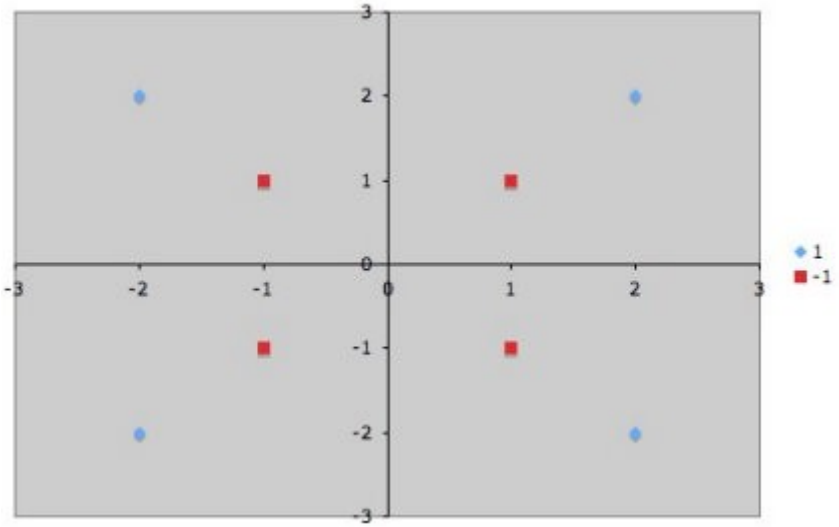
Exemplo (7)

- (1.5) mapear para (1.5, 2.25)
  - -3 . 1.5 + 2.25 + 1 = - 1.15 [-1]
- (-1) mapear para (-1,1)
  - -3 . -1 + 1 + 1 = 5 [+1]
- (4) mapear para (4,16)
  - Exercício

Segundo exemplo

- Como separar as duas classes com apenas uma reta?

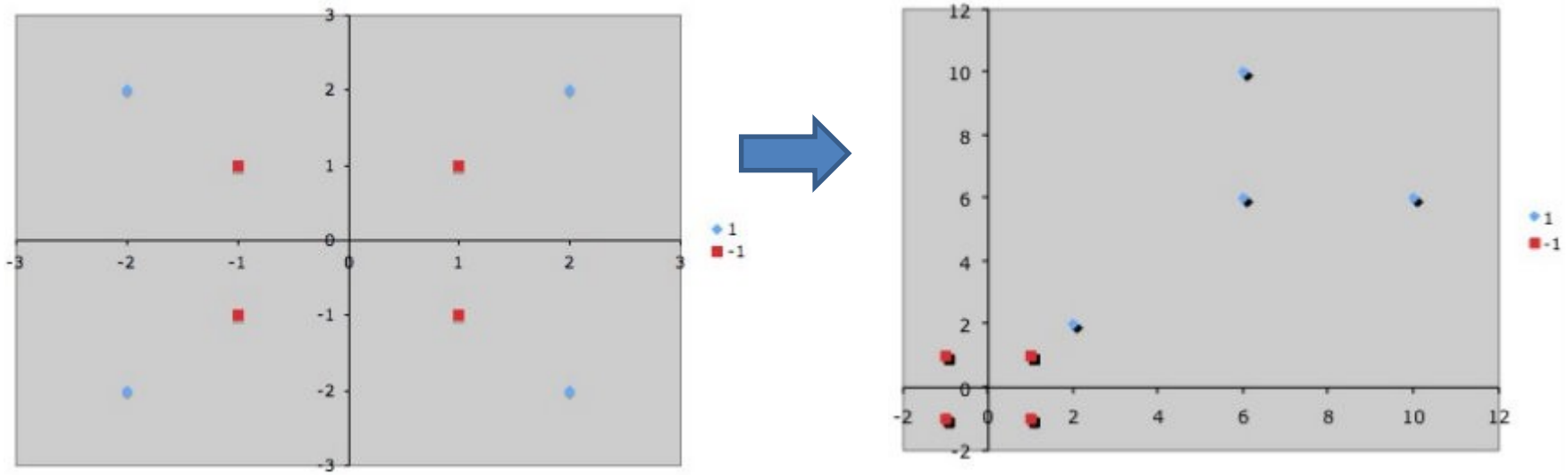
X_1	X_2	Class
1	1	-1
-1	1	-1
1	-1	-1
-1	-1	-1
2	2	+1
-2	2	+1
2	-2	+1
-2	-2	+1



Segundo exemplo (2)

$$\Phi(x_1, x_2) = \begin{cases} (4-x_2+|x_1-x_2|,\ 4-x_1+|x_1-x_2|), & \sqrt{(x_1^2 + x_2^2)} > 2 \\ (x_1, x_2) \end{cases}$$

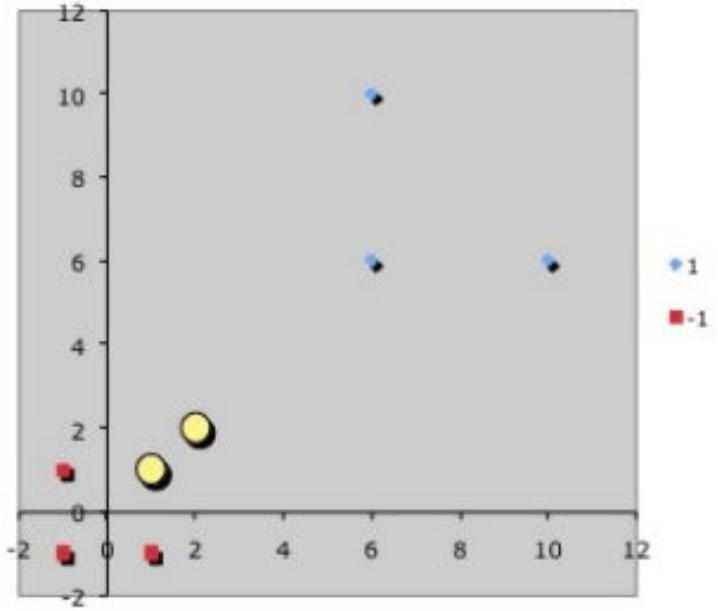
- Esta função mantém o problema bidimensional



Segundo exemplo (3)

- Vetores de Suporte

x_1	x_2	Class
1	**1**	**-1**
-1	1	-1
1	-1	-1
-1	-1	-1
2	**2**	**+1**
6	6	+1
10	6	+1
6	10	+1



Segundo exemplo (4)

- Vetores de Suporte

x_1	x_2	Class
1	**1**	**-1**
2	**2**	**+1**

- Como só temos dois pontos, vamos cair em um sistema com duas equações e três incógnitas
- $w_1 + w_2 + b = -1$
  $2w_1 + 2w_2 + b = +1$

Segundo exemplo (5)

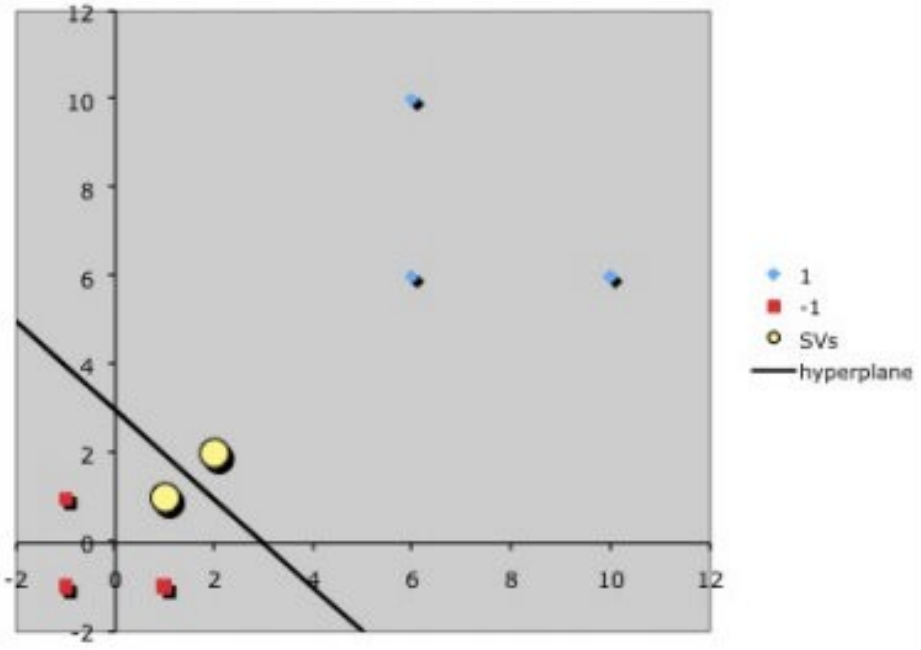
- Contudo, podemos verificar que a distância euclidiana entre os pontos é $\sqrt{2}$
- Logo, a distância entre os pontos e o hiperplano deve ser $\sqrt{2}/2$
- Vamos recorrer a fórmula para calcular a distância entre reta e ponto:
- $d(p, h) = |w_1x_1 + w_2x_2 + b| / \sqrt{(w_1^2 + w_2^2)}$
- Vamos usar (1, 1) como ponto de referência e, **removendo o módulo**, temos $-\sqrt{2}/2$ (ponto de referência é da classe negativa)

Segundo exemplo (6)

- Após os devidos ajustes:
  - $w_1 + w_2 + b = -1$
    $2w_1 + 2w_2 + b = +1$
    $(w_1 + w_2 + b)/\sqrt{(w_1^2 + w_2^2)} = -\sqrt{2}/2$
  - Resolva este sistema

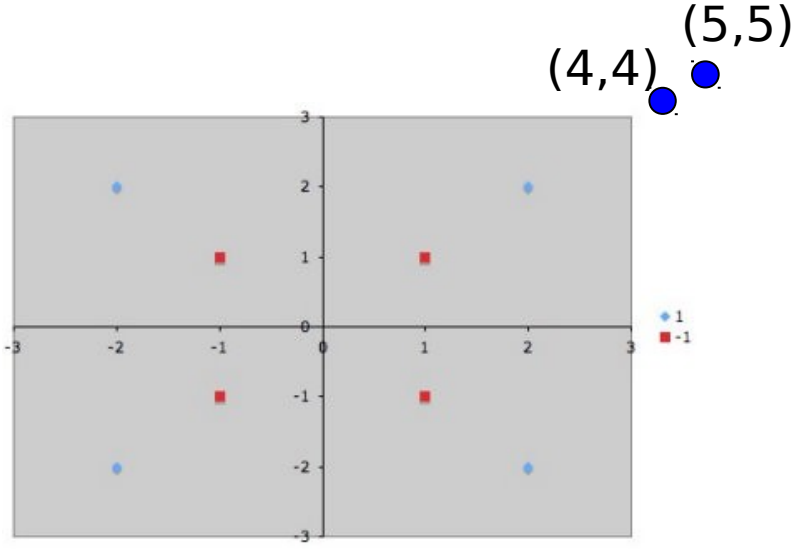
Segundo exemplo (7)

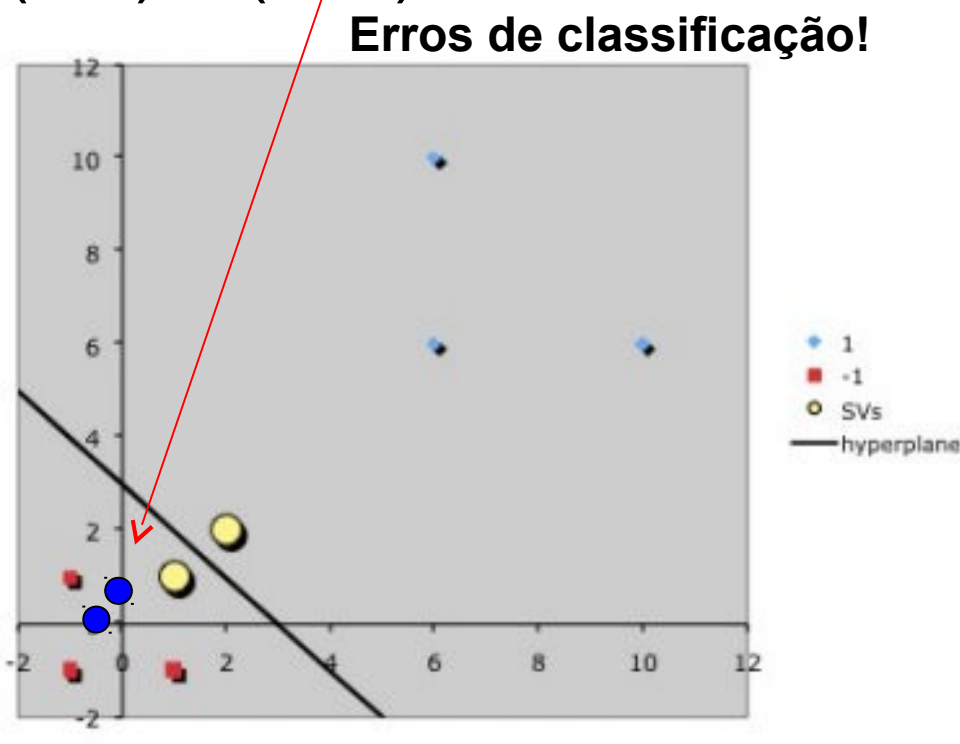
- h_0: (1,1) . (x_1, x_2) -3 = 0
  $x_1 + x_2 - 3 = 0$



Segundo exemplo (7)

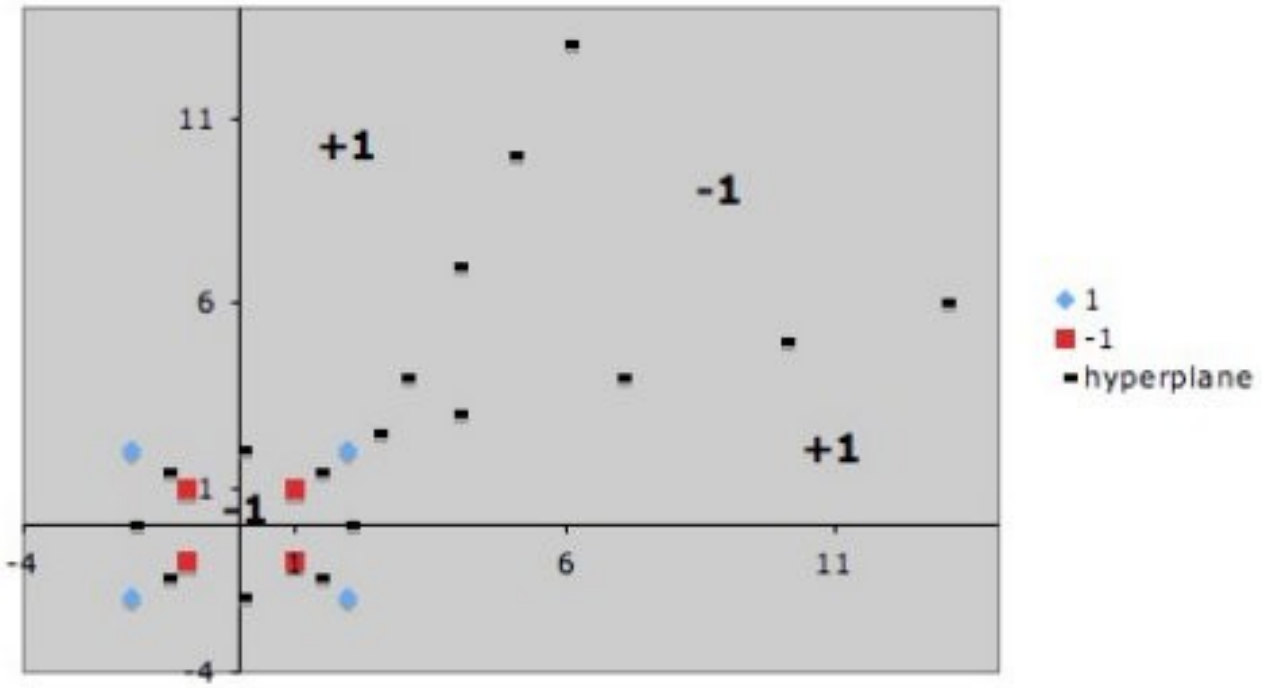
- Esta função realmente separa o espaço original de forma linear?
  Dados de teste (5,5) $\Phi(5,5) = (-1, -1)$
  Dados de teste (4,4) $\Phi(4,4) = (0, 0)$





Segundo exemplo (9)

- Função de mapeamento não é ideal



Funções Kernel

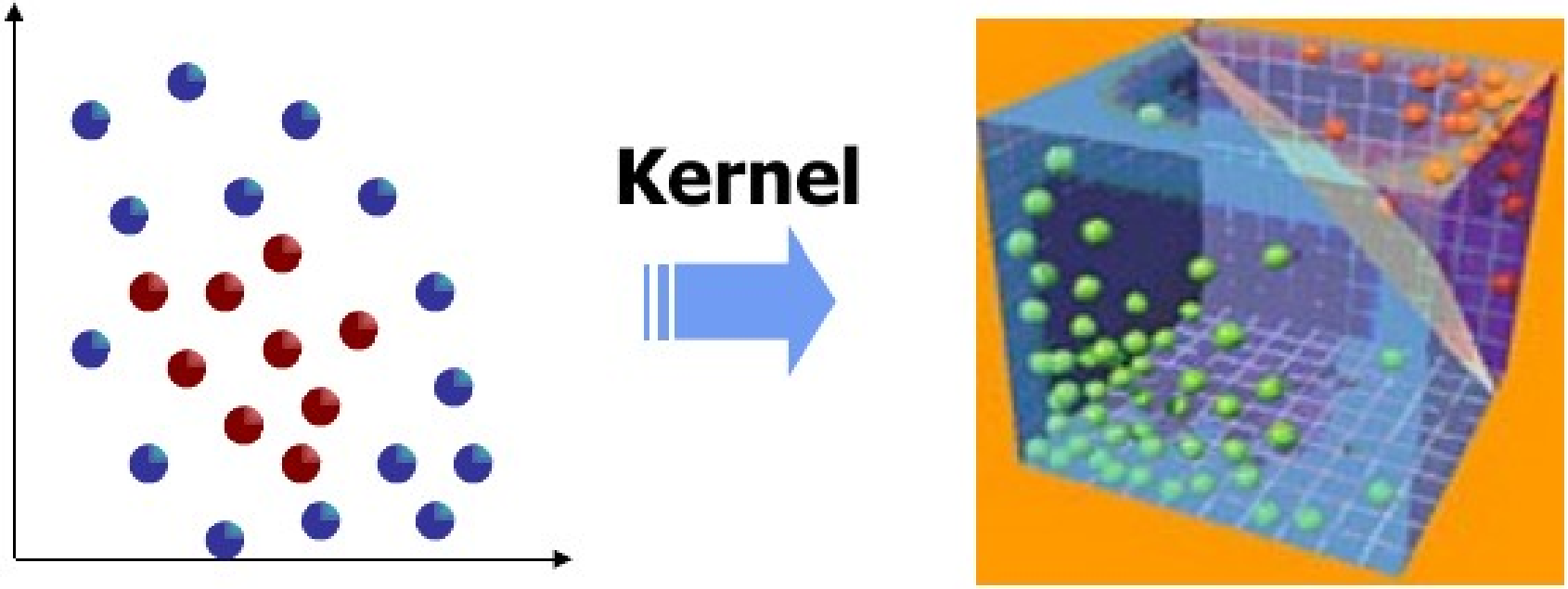
- Informalmente, são funções em que a ordem dos parâmetros não altera o resultado.
- Exemplo $R^2 \rightarrow R^1$:
  - $k(x, y) = x^2 + 2xy + y^2$
  - $k(2, 3) = k(3, 2) = 25$

Funções Kernel (2)

- Exemplo $R^2 \rightarrow R^3$:
  - $f(x, y) = (x^2, 2xy, y^2)$
  - $f(2, 3) = (4, 12, 9)$
  - Obs: não é kernel, só foi inspirada em uma por simplicidade

Funções Kernel (3)

- Usar uma função Kernel adequada é equivalente a mapear o problema em uma dimensão maior, na qual os dados são linearmente separáveis



Avançado: kernel trick

- **Kernel trick** (truque kernel)
  - Nossa implementação é muito simples, baseada em sistemas de equação
  - Implementações reais de SVM usam produtos escalares
  - Kernel trick é uma maneira de usar usando produtos escalares para subir muitas dimensões sem precisar fazer o mapeamento explicitamente

Avançado: kernel trick (2)

- Suponha x = (a, b) no R^2 e que vamos mapear esses dados para o R^3
- Temos: $f(a, b) = (a^2, ab\sqrt{2}, b^2)$
- Exemplo:
  - $p_1 = (1, 2) \rightarrow f(p_1) = (1, 2\sqrt{2}, 4)$
  - $p_2 = (3, 4) \rightarrow f(p_1) = (9, 12\sqrt{2}, 16)$

Avançado: kernel trick (2)

- $p_1.p_2 = (1, 2) . (3, 4)$
  $= 3 + 8 = 11$
- $f(p_1).f(p2) = (1, 2\sqrt{2}, 4) . (9, 12\sqrt{2}, 16)$
  $= 9 + 48 + 64 = 121$
- Note que $(p_1.p_2)^2 = f(p_1).f(p_2)$

Avançado: kernel trick (3)

- Solução simples: usamos *f* para mapear *x* para R^3 e calculamos produto escalar: f(x) . f(x)
- Solução otimizada: usamos um kernel equivalente a f. No exemplo: $k = (x . x)^2$
- **Conseguimos um produto escalar do R^3 sem precisar ir para o R^3!**

Funções Kernel

- Mais usadas
  - Polinomial
  - Gaussiana
  - Sigmoid
- Sempre aumentam o número de dimensões
  - Normalmente, aumentam bastante!

Créditos

- Eduardo Borges e Gabriel Simões da UFRGS