Support Vector Machines

Uma breve introdução...

Eduardo Corrêa Gonçalves

16 de Abril 2013

Tópicos da Aula

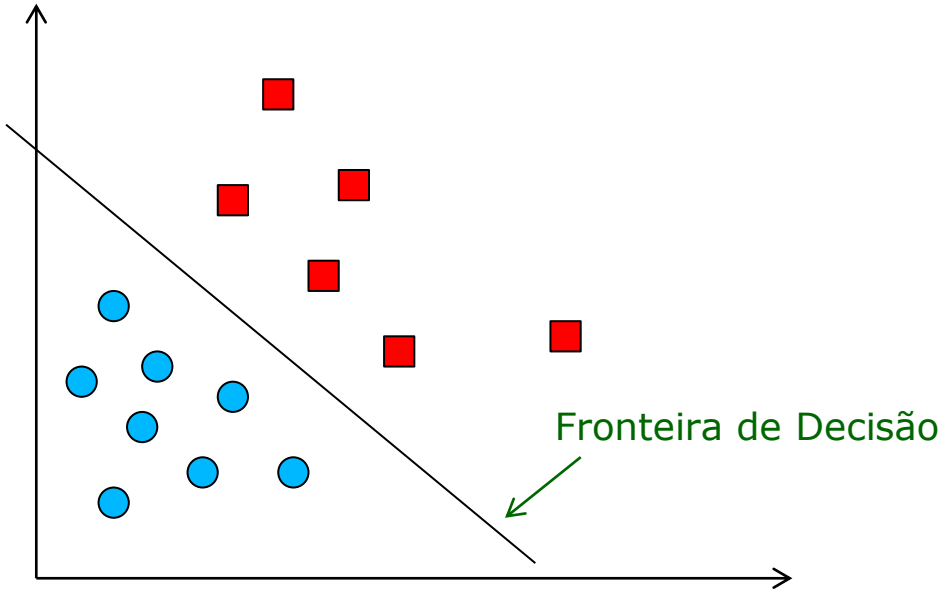
- **Introdução**
  - SVM: Características
  - SVM: Intuição
- **Caso 1: Classificação de Dados Linearmente Separáveis**
  - Apresentação do Exemplo
  - Dados Linearmente Separáveis
  - Hiperplano, Margem e Margem Máxima
  - Treinamento do SVM
  - Classificação com o SVM
- **Caso 2: Classificação de Dados Não Lineares**
  - Método usado pelo SVM
  - Funções Kernel
- **Comentários Finais**
- **Bibliografia**

Introdução (1/4)

- **SVM – Características**
  - 😐 Desenvolvido, em grande parte, na AT&T Bell
  - 😐 Primeiro paper publicado em 1992.
    - Porém, possui raízes na "Teoria do Aprendizado Estatístico" (1960s)
  - ☺ Grande acurácia em vários domínios.
  - ☺ Menos sujeito ao problema de *overfitting* (superajuste)
  - ☺ Pode ser utilizado para a classificação de dados lineares e não-lineares.
  - ☹ Tempo de treinamento alto.
  - 😐 Possui fundamentos teóricos sofisticados.

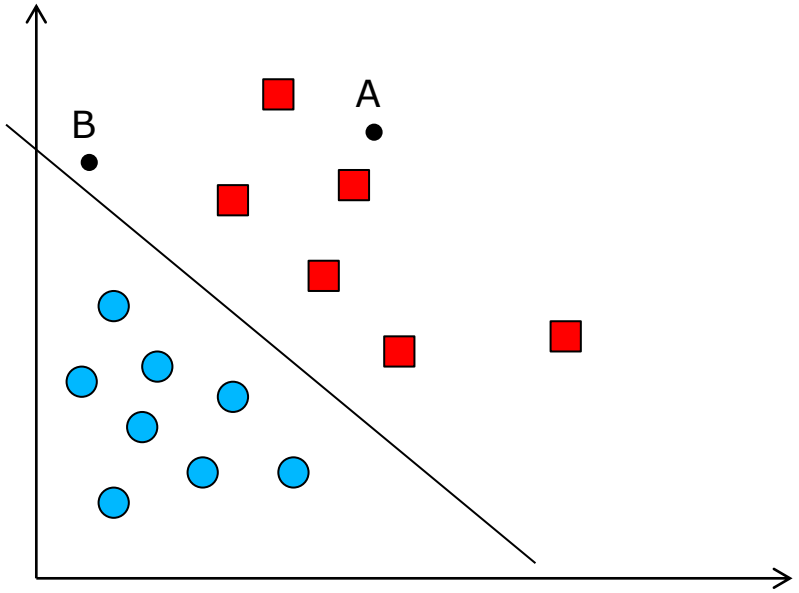
Introdução (2/4)

- **SVM: Intuição** [Ng, 2012]
  - Tuplas de classe ■ e ●
  - Uma fronteira de decisão gerada por um algoritmo



Introdução(3/4)

- **SVM: Intuição** [Ng, 2012]
  - **Confiança da Predição**
  - Considere os pontos A e B

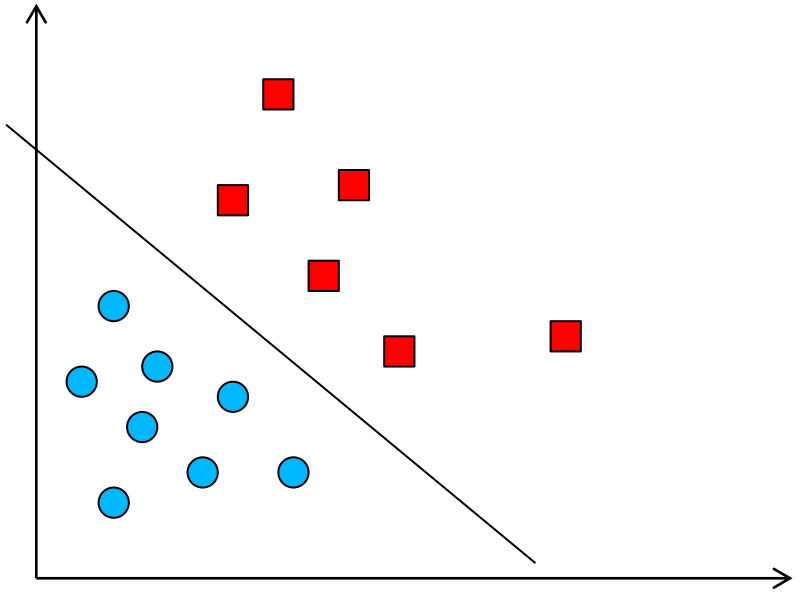


$y \in \{ \bigcirc, \blacksquare \}$

- **A** está muito distante da fronteira.
  - Se eu "arriscasse" uma predição para A, eu diria que é ■ com muita confiança.
- **B** está muito próximo da fronteira.
  - Se essa fronteira se modificar um pouco, talvez B deixe de ser ■.
- **Ou seja**: intuitivamente, temos muito mais confiança na predição de A do que de B.

Introdução (4/4)

- **SVM: Intuição** [Ng, 2012]
  - O SVM parte da seguinte ideia básica:
    - A partir de uma base de dados de treinamento, é interessante encontrar uma fronteira de decisão que nos permita fazer todas as predições **corretamente** e com **grande confiança**.



Distante da fronteira de decisão

Caso 1: Dados Lineares (1/21)

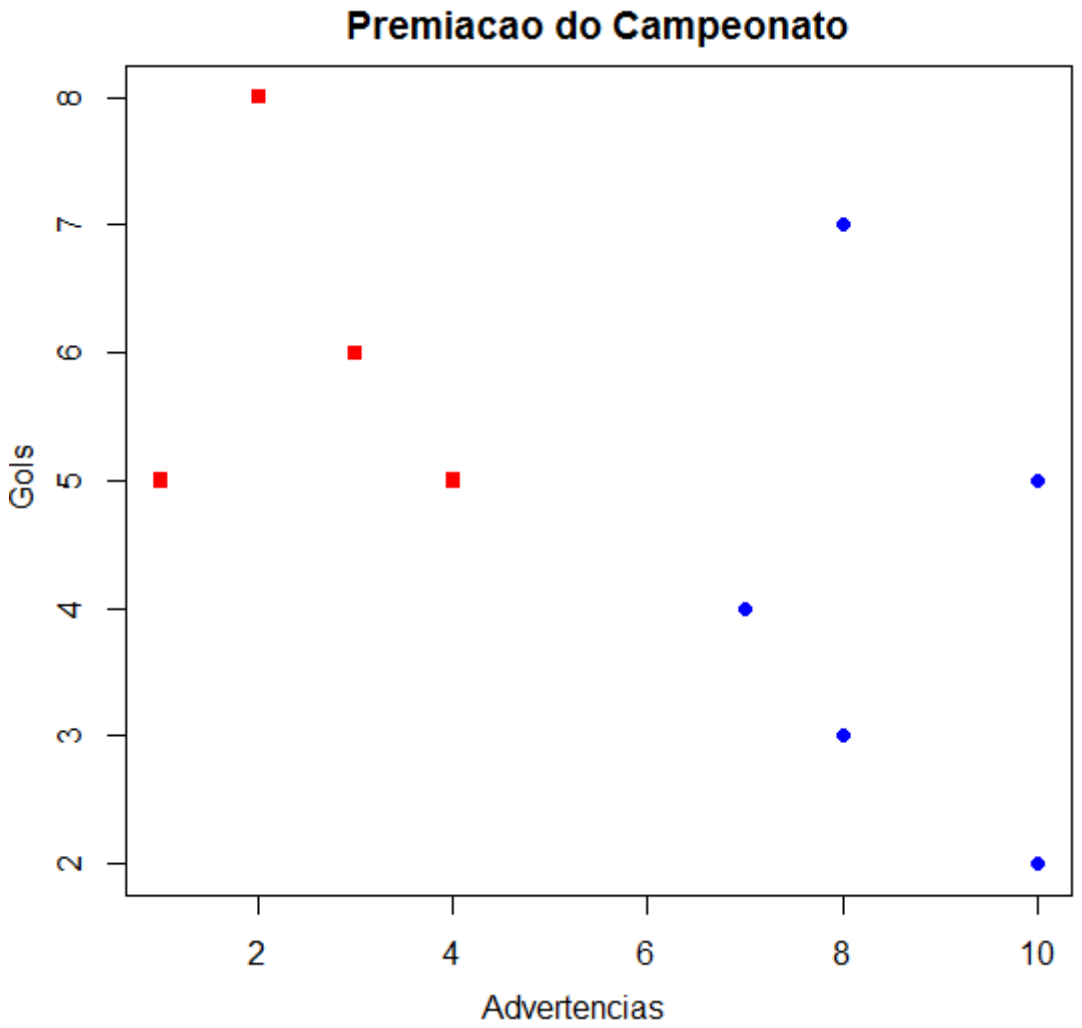
- **Base de Dados de Treinamento (*D*)**
  - Premiação de um Campeonato de Futebol

advertências (X1)	gols (X2)	classe (y)
1	5	1
4	5	1
2	8	1
3	6	1
8	3	-1
7	4	-1
10	2	-1
10	5	-1
8	7	-1

- 2 atributos de entrada: X1 e X2
- N = 9
- $y \in \{-1, +1\}$
  - *O SVM usa essa notação para y, ao invés de {0,1}*

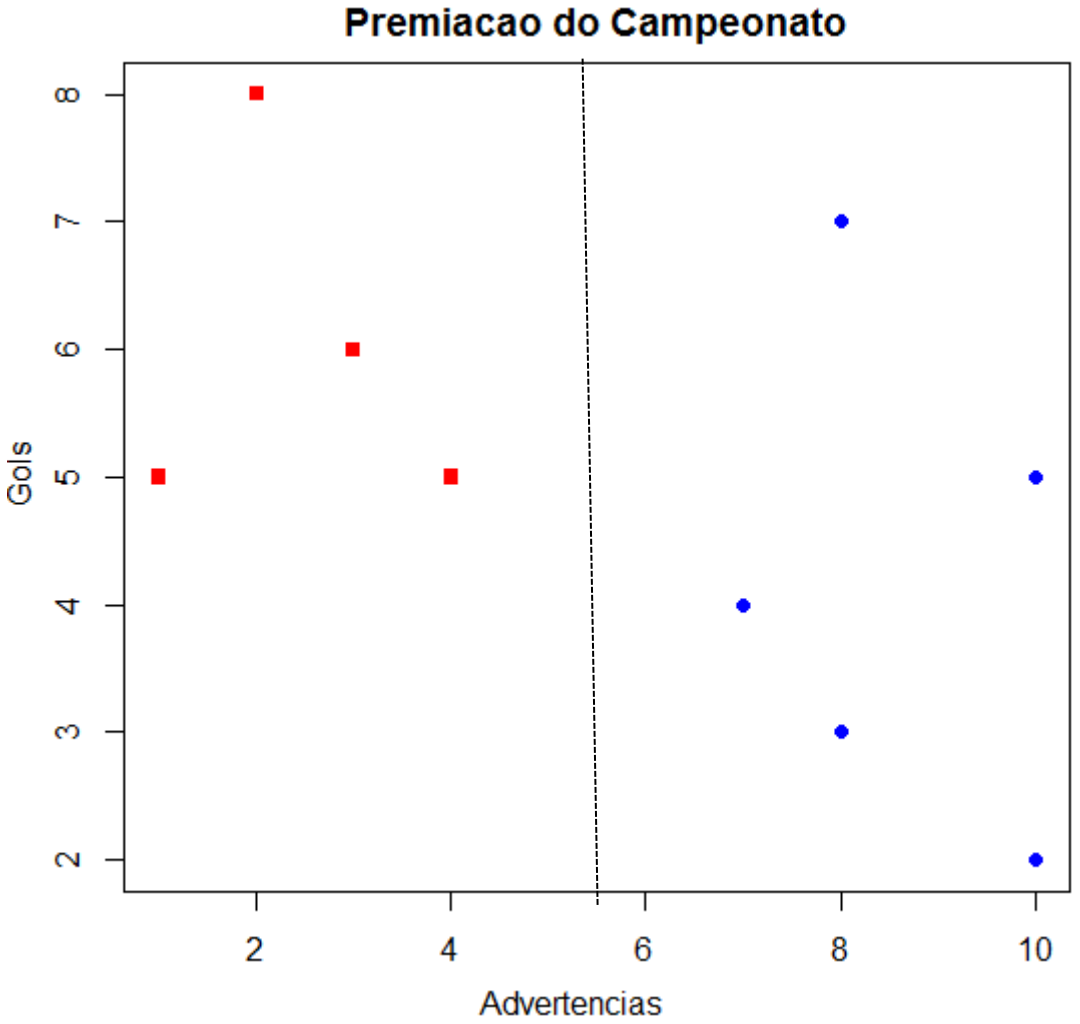
Caso 1: Dados Lineares (2/21)

- **Diagrama de Dispersão do BD**
  - Os exemplos pertencem a duas classes: (■ = +1, ● = -1)



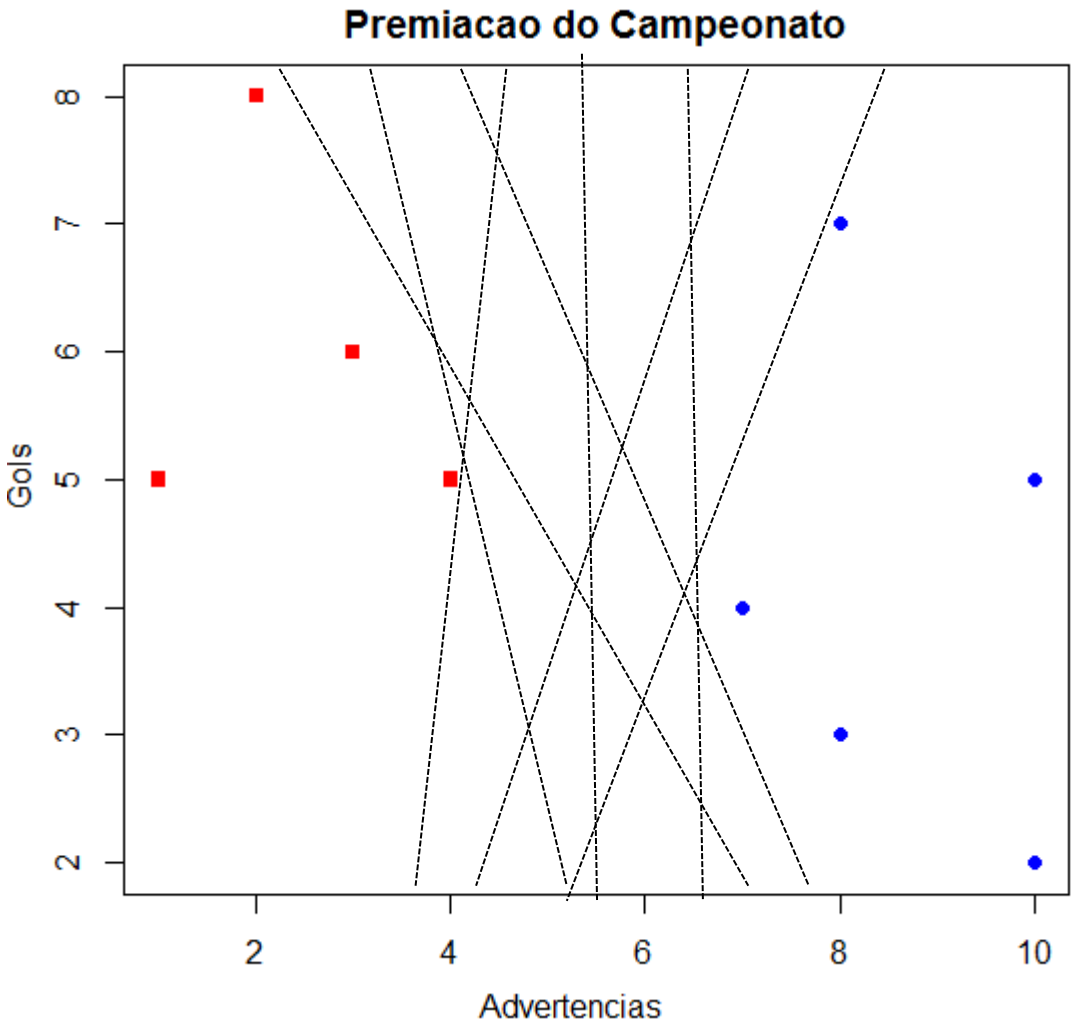
Caso 1: Dados Lineares (3/21)

- **Hiperplano Separador (1/2)**
- Os dados dão **linearmente separáveis**.
    - **Linha reta** (**hiperplano**) pode ser desenhada de modo que todas as tuplas da classe **+1** fiquem de um lado e as de classe **-1** do outro.



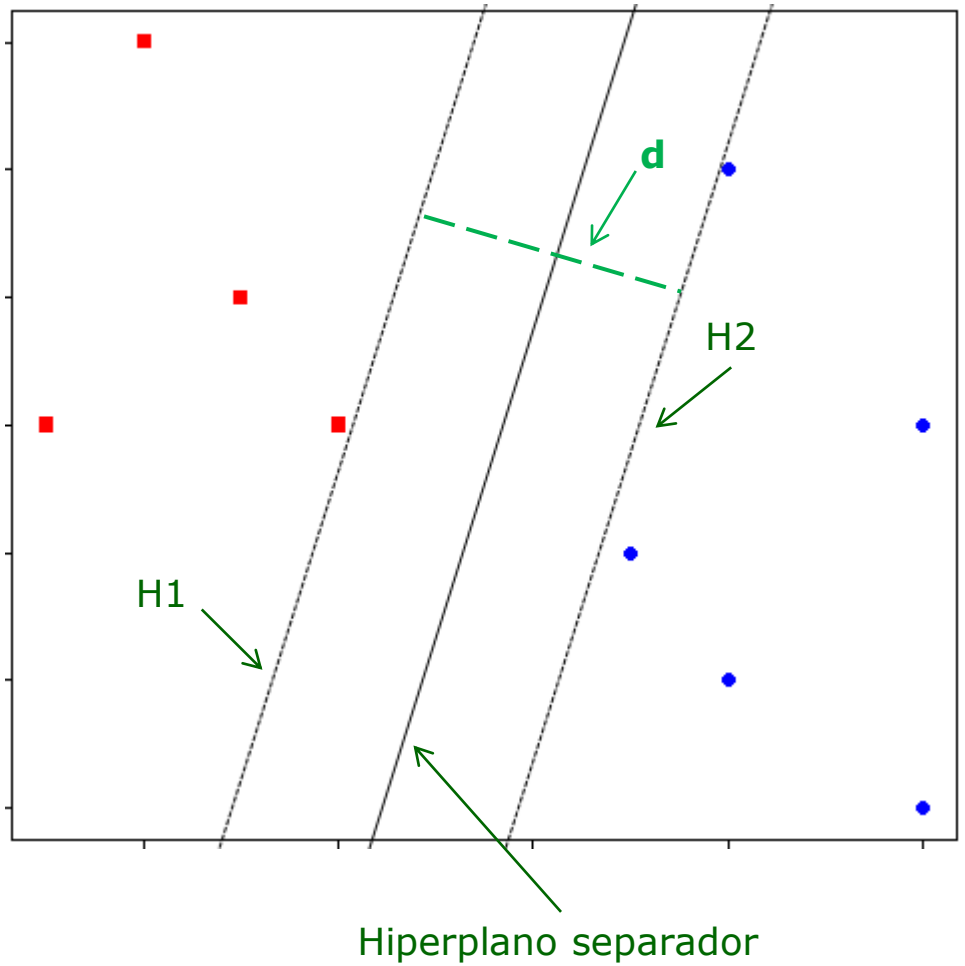
Caso 1: Dados Lineares (4/21)

- **Hiperplano Separador (2/2)**
- De fato, **infinitos** hiperplanos separadores poderiam ser desenhados com erro de treinamento = 0
    - Porém, não há garantias de que a mesma performance se repetirá para novos exemplos.
    - O SVM deverá escolher um deles, baseado em quão bem funcionará para novos exemplos.



Caso 1: Dados Lineares (5/21)

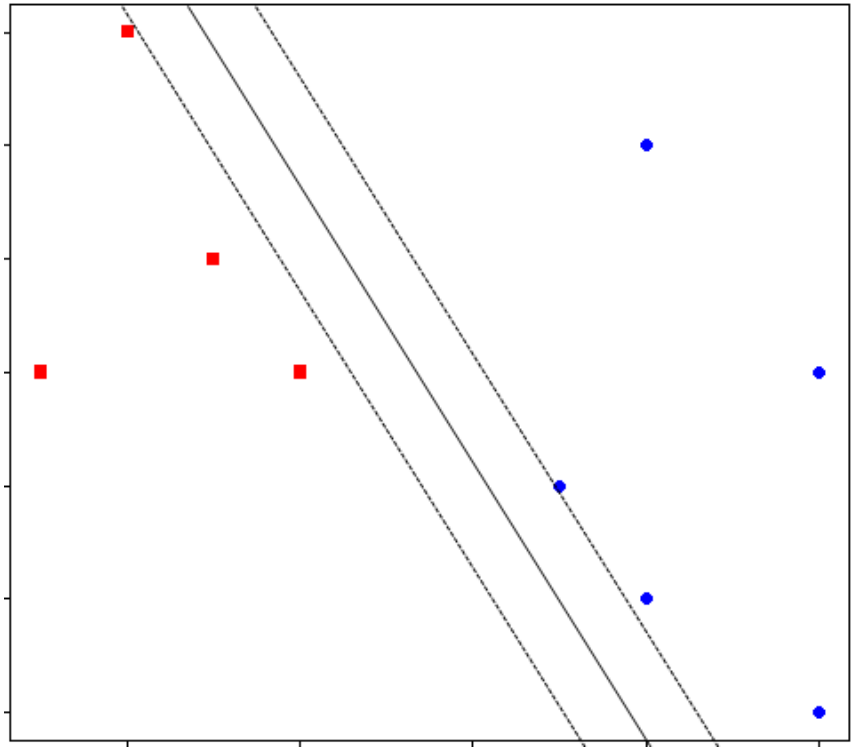
- **Margem (1/2)**
  - O SVM toma a sua decisão baseado no conceito de margem.



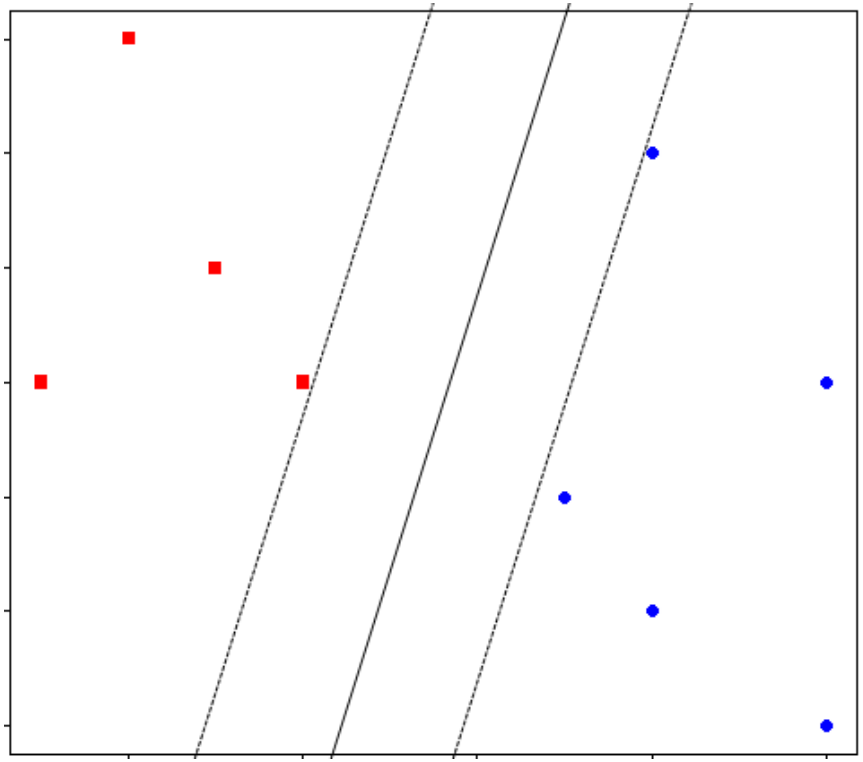
- **H1** é obtido movendo um hiperplano paralelo ao hiperplano separador até tocar no ■ mais próximo.
- De maneira análoga, obtém-se **H2** para ● .
- **margem** = distância **d** entre H1 e H2.

Caso 1: Dados Lineares (6/21)

- **Margem (2/2)**
  - Recorde que existem infinitos hiperplanos separadores (diferentes margens)
  - No exemplo abaixo, ambos os hiperplanos classificam corretamente todas as tuplas. Qual escolher?



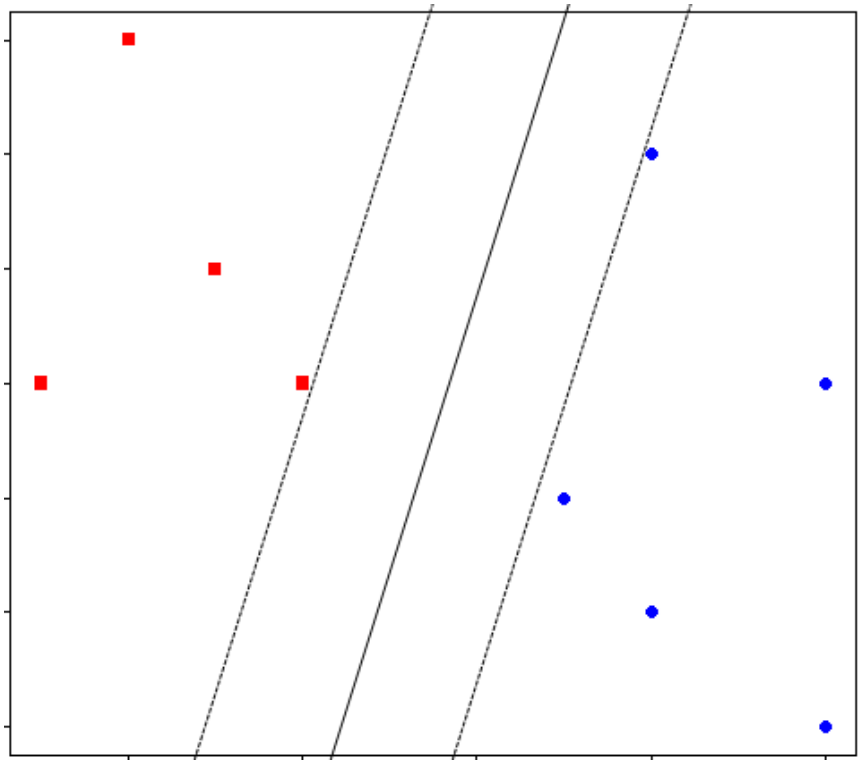
Margem pequena



Margem grande

Caso 1: Dados Lineares (7/21)

- **Hiperplano de Margem Máxima** (*Maximum Margin Hyperplane* **- MMH**)
  - O SVM procura o hiperplano de margem máxima.



Margem máxima (**MMH**)

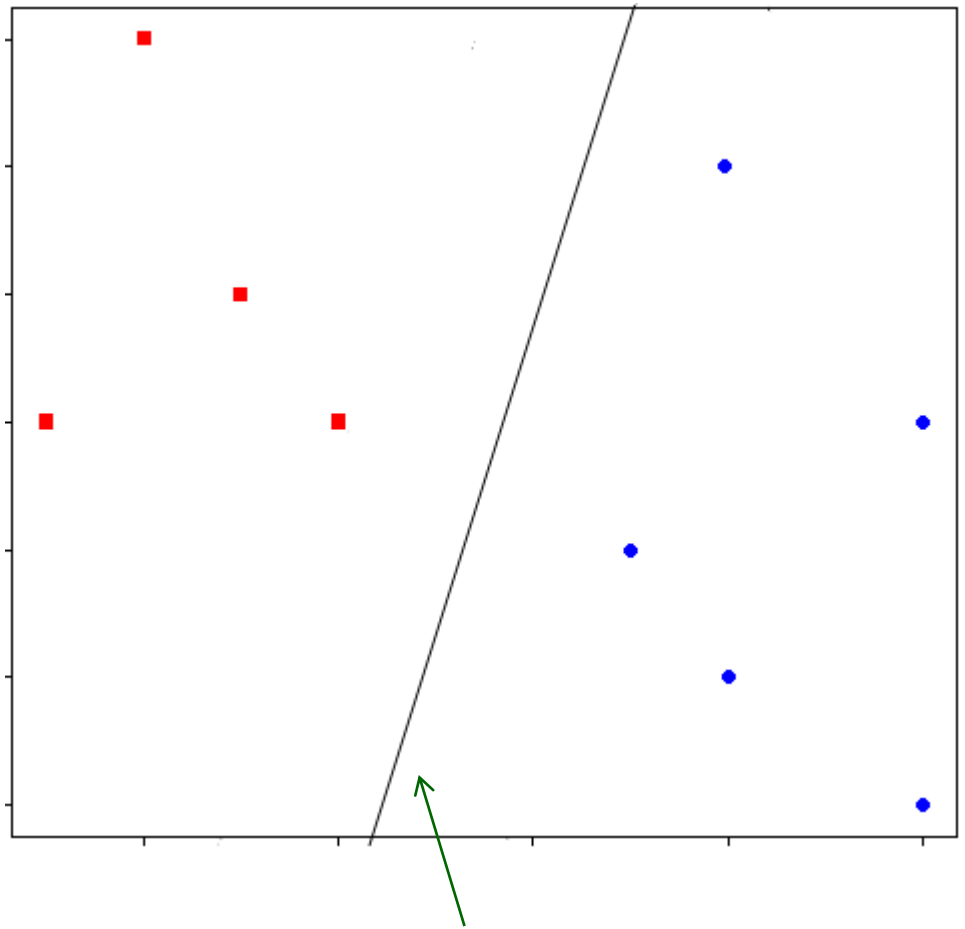
- Da **intuição** apresentada no início da aula:
  - Confiança da predição é maior se a margem é maior.
  - Se margem é pequena, qualquer perturbação na fronteira de decisão pode ter impacto significante na classificação.
    - Fronteiras com margem pequena são mais suscetíveis ao superajuste.

Caso 1: Dados Lineares (8/21)

- **Porque MMH?**
  - Intuição.
  - Resultados Empíricos
  - Mais formalmente: Teoria do Aprendizado Estatístico
    - Structural Risk Minimization
    - Referências: [*Burges 1998*], [*Smola & Scholkopf 1998*]

Caso 1: Dados Lineares (9/21)

- **Classificador SVM Linear**
  - Busca pelo hiperplano de margem máxima.



- **Equação do Hiperplano**
  - Um hiperplano separador pode ser escrito como:

    $WX + b = 0$

  - Onde:
    - **W** = {w1, w2, ..., wn} é um vetor de pesos
    - **b** é um escalar

(qualquer exemplo que satisfaça a equação cairá aqui)

Caso 1: Dados Lineares (10/21)

- **Equação do Hiperplano**
  - Em nosso caso temos dois atributos:

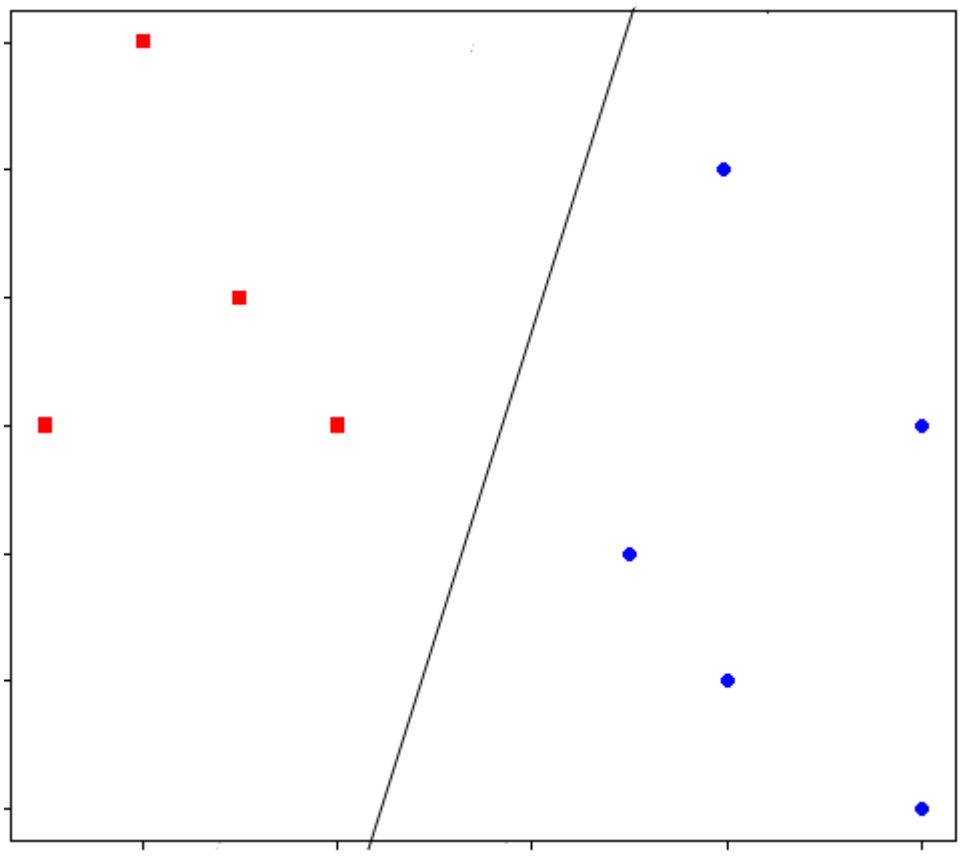
    $w_1x_1 + w_2x_2 + b = 0$

  - Qualquer ponto acima da fronteira satisfaz:

    $w_1x_1 + w_2x_2 + b > k, \quad k > 0$

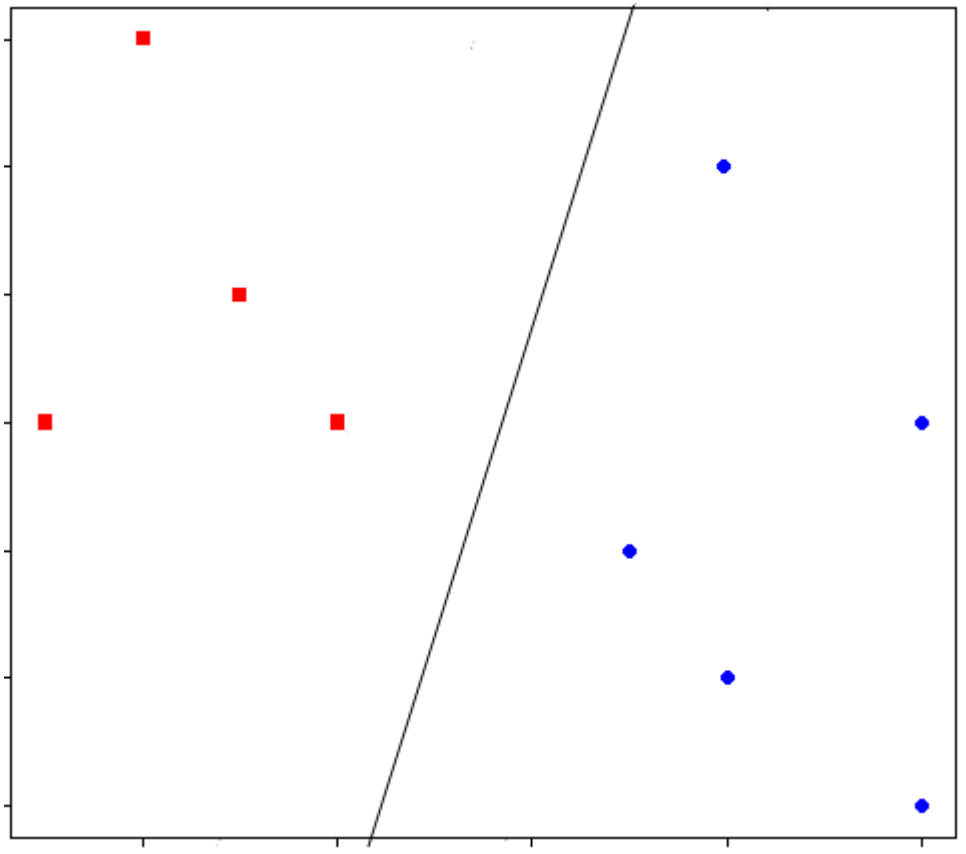
  - Qualquer ponto à abaixo da fronteira satisfaz:

    $w_1x_1 + w_2x_2 + b < k', \quad k'<0$



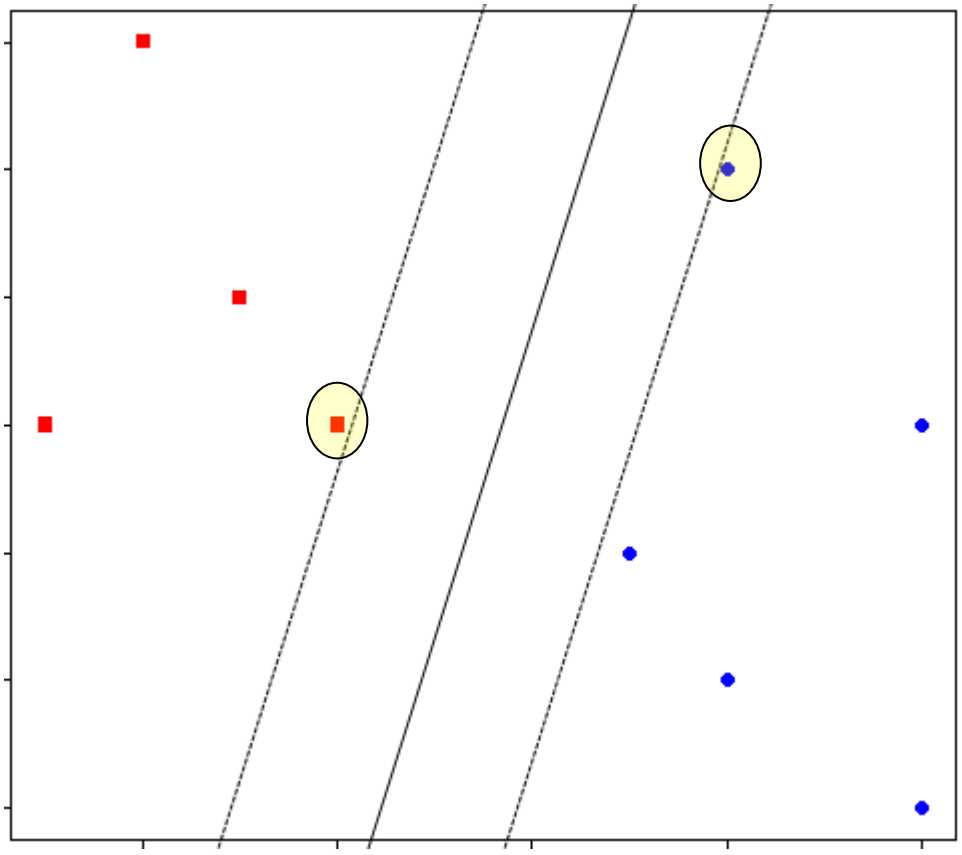
Caso 1: Dados Lineares (11/21)

- **Equação do Hiperplano**
  - Como:
  - ■ = label + 1; e
  - ● = label -1.
  - Podemos prever a classe y de qualquer exemplo de teste z fazendo:
  - $y = 1$, se $wz + b \geq 0$
  - $y = -1$, se $wz + b < 0$



Caso 1: Dados Lineares (12/21)

- **Support Vectors (1/2)**
  - Considere o ■ e a ● destacados
    - Eles são os pontos mais próximos da fronteira de decisão.
    - Eles são chamados de **support vectors (SV's)**.



  - ■ satisfaz:

$w_1x_1 + w_2x_2 + b = k \;,\; k > 0$

  - ● satisfaz:

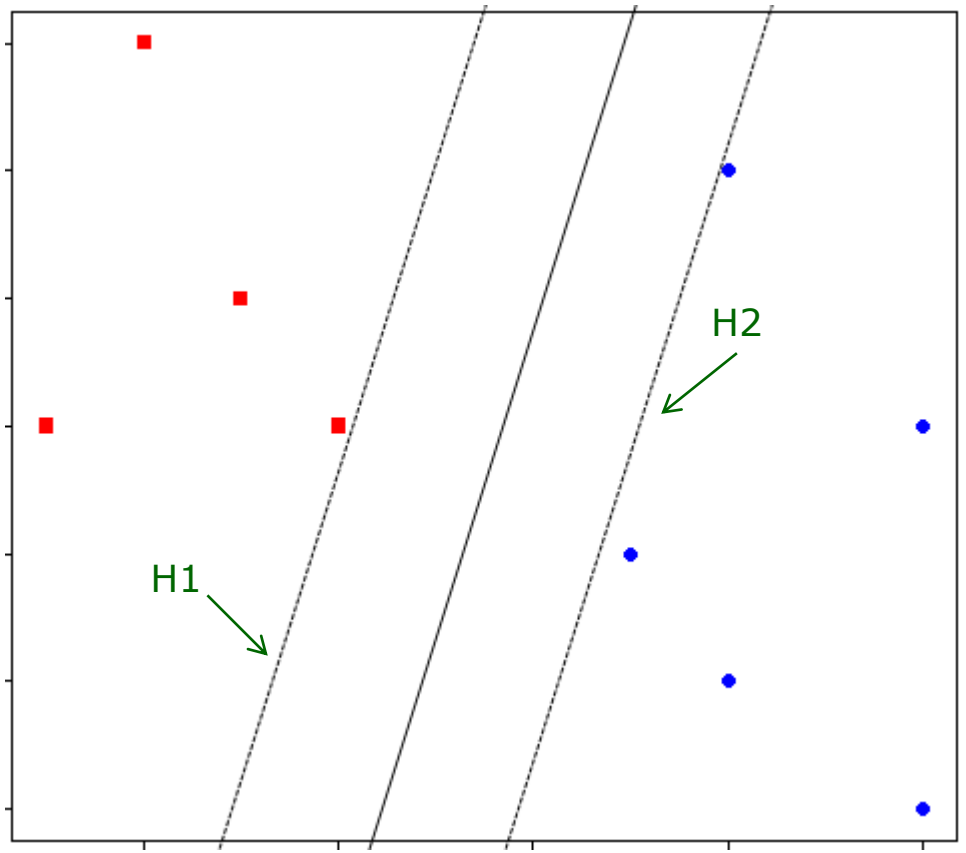
$w_1x_1 + w_2x_2 + b = k' \;,\; k' < 0$

Caso 1: Dados Lineares (13/21)

- **Support Vectors (2/2)**
  - Podemos reescalar os parâmetros w e b de forma que 2 hiperplanos paralelos H1 e H2 sejam expressos na forma:

    H1: $w1x1 + w2x2 + b \geq 1$

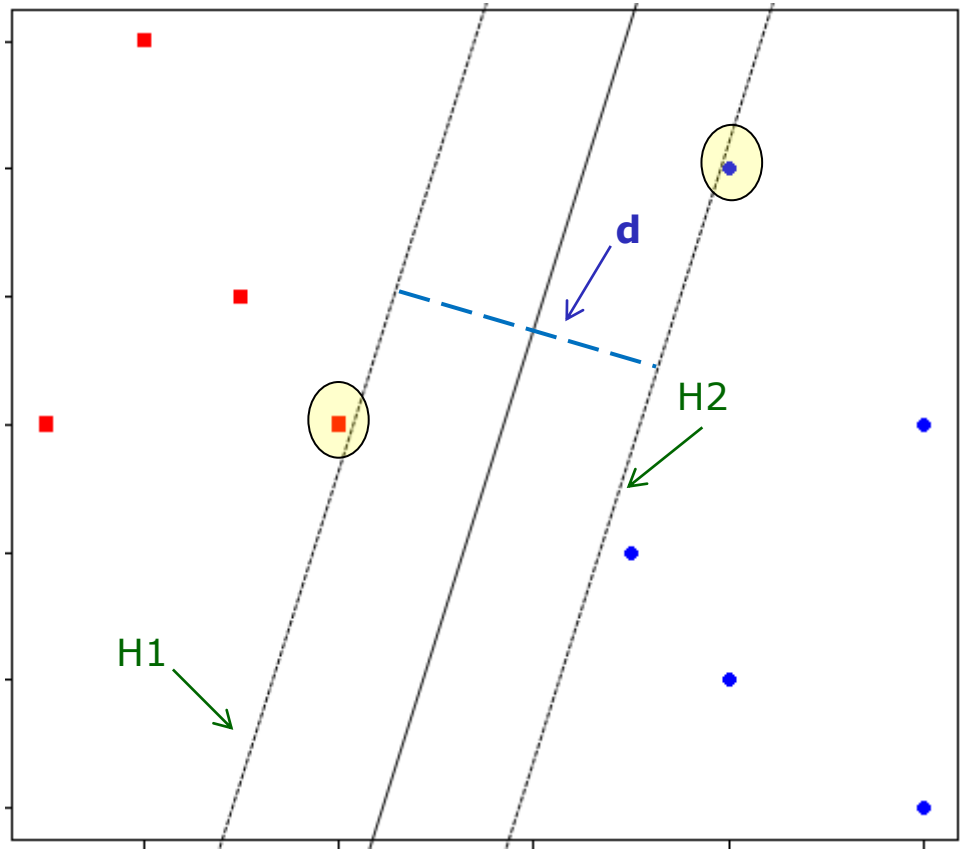
    H2: $w1x1 + w2x2 + b \leq -1$



- Qualquer tupla sobre ou acima de H1 pertence à classe +1
- Qualquer tupla sobre ou abaixo de H2 pertence à classe -1.

Caso 1: Dados Lineares (14/21)

- **Margem Definida pelos SV's**
- A margem (d) é dada pela distância entre H1 e H2 (**cálculo vetorial**)
  - A distância da fronteira para qualquer ponto em H1 é 1/||W||.
  - Idem para H2.
- Portanto o **tamanho** margem máxima é:
  - d = 2 / ||W||

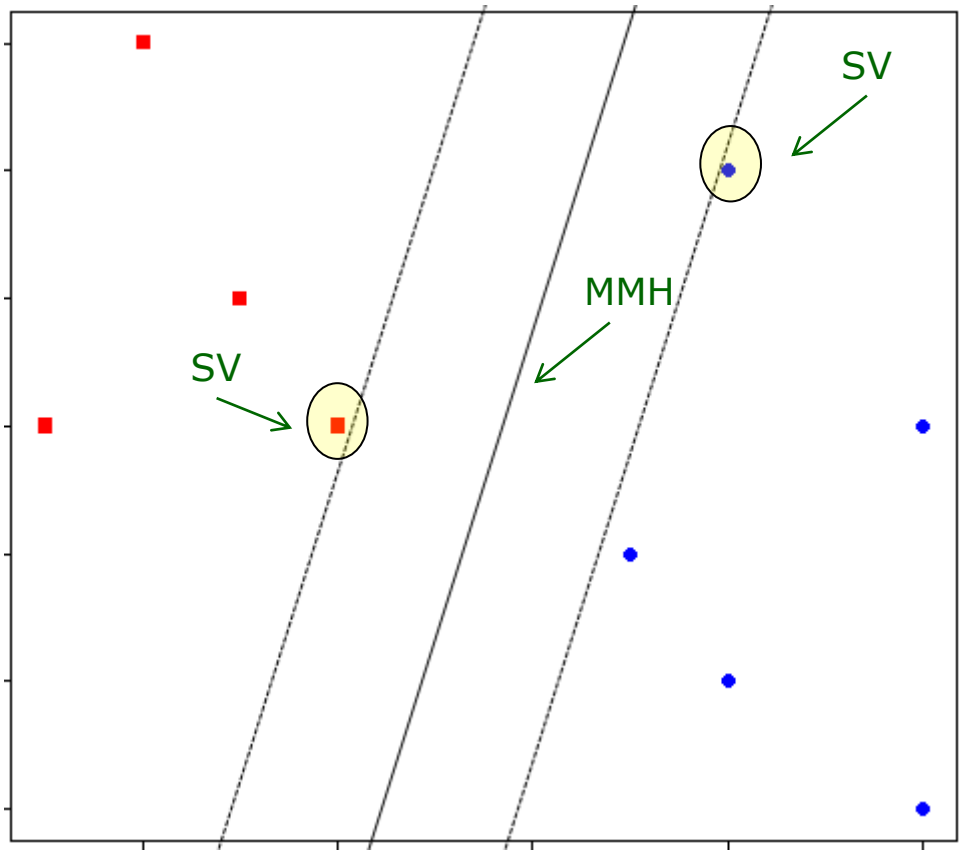


- ||W|| = norma do vetor W.

$$\| W \| = \sqrt{W \times W} = \sqrt{w1^2 + w2^2 + ... + wn^2}$$

Caso 1: Dados Lineares (15/21)

- **Support Vector Machines**
- Qual o objetivo do SVM?
    - Encontrar os SV's. Por consequência, encontrar o MMH



Caso 1: Dados Lineares (16/21)

- **SVM: treinando o modelo (1/3)**
  - Treinar o SVM, na verdade, envolve estimar os parâmetros w e b do modelo.
  - Eles devem ser escolhidos de modo que as seguintes duas condições sejam satisfeitas:

    $wx_i + b \geq 1$ se $y_i = 1$

    $wx_i + b \leq -1$ se $y_i = -1$

  - Relembrando que estamos no caso “linear”:
    - As condições acima impõem como requisito que todas as instâncias de treino da classe y=+1 estejam acima do hiperplano wx + b = 1.
    - De maneira equivalente, para y=-1, abaixo do hiperplano wx+b = -1.

Caso 1: Dados Lineares (17/21)

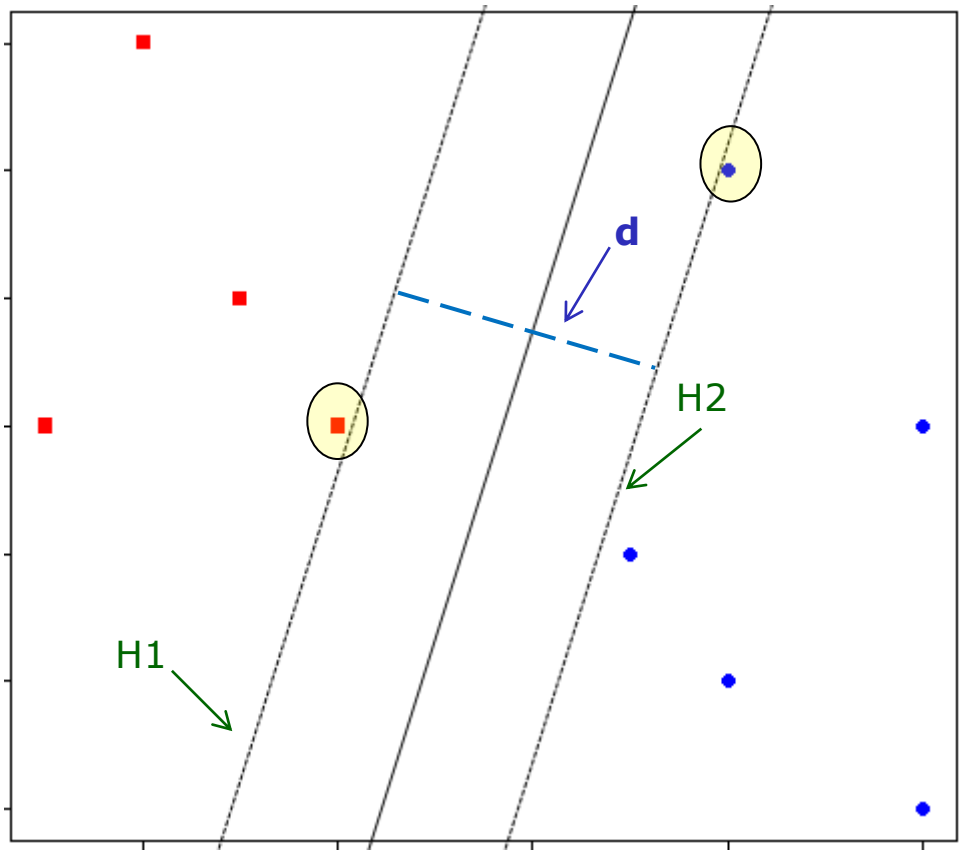
- **SVM: treinando o modelo (2/3)**
  - As inequações:

    $wx_i + b \geq 1$ se $y_i = 1$

    $wx_i + b \leq -1$ se $y_i = -1$
  - Podem ser reescritas como:

    $y_i(wx_i + b) \geq 1$
  - Seria fácil resolver, mas o SVM impõe um requisito adicional:
    - A margem deve ser **máxima**!
    - Maximizar a margem equivale a minimizar a função objetivo:

      $f(w) = ||w||^2 / 2$



Caso 1: Dados Lineares (18/21)

- **SVM: treinando o modelo (3/3)**
  - Então, temos o problema definido:

Minimizar w $\quad \mathbf{||w||^2 / 2}$

s.a. $\mathbf{y_i(wx_i + b) \geq 1}$

  - A função objetivo é quadrática e as restrições lineares em w e b:
    - PROBLEMA DE OTIMIZAÇÃO CONVEXO
      - Resolvido pelo método padrão: "multiplicador de Lagrange" (consulte [Ng, 2012], [Tan et al. 2006])

Caso 1: Dados Lineares (19/21)

- **Classificando Novas Tuplas (1/2)**
- Baseado na formulação Lagrangiana, a MMH pode ser reescrita como a fronteira de decisão:

$$d(z) = \sum_{i=1}^{l} yi \times \alpha i \times xi \times z + bo$$

- A partir desta fórmula podemos classificar novos objetos.
  - l: número de SV's
    - Para dados linearmente separáveis, os SV's são um subconjunto das tuplas de treino.
    - Observe que, no somatório, os **SV's são as únicas tuplas** da base de treinamento que são **levadas em consideração**.
  - yi: rótulo de classe do support SV xi
  - z: tupla de teste
  - αi e bo: parâmetros numéricos determinados pelo algoritmo SVM
    - αi são multiplicadores Lagrangianos

Caso 1: Dados Lineares (20/21)

- **Classificando Novas Tuplas (2/2)**
- Dada uma tupla de teste z , basta "pluga-la" na equação e observar o **sinal** do resultado.
    - Ele nos diz em que lado do hiperplano a tupla de teste "cai".
        - Se positivo, classe é +1
        - Se negativo, classe é -1

$$d(z) = \sum_{i=1}^{l} yi \times \alpha i \times xi \times z + bo$$

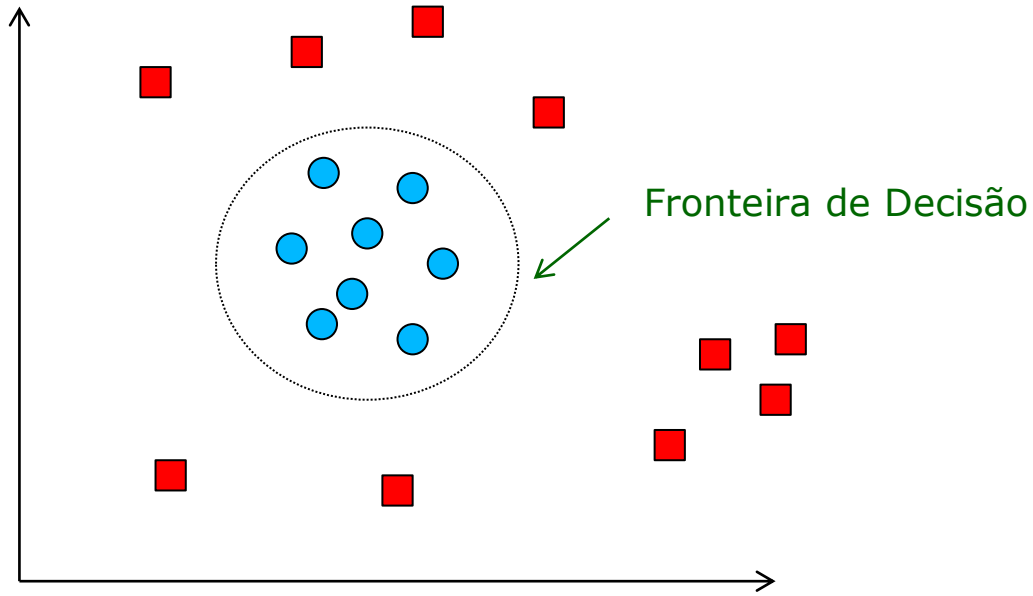
Caso 1: Dados Lineares (21/21)

- **Observações Finais**
- SV's são as tuplas de treinamento essenciais
    - Se todas as outras tuplas fossem removidas e o treinamento repetido, o mesmo hiperplano separador seria encontrado.
    - Com isso, o SVM consegue definir um modelo de classificação baseado em pouquíssimos elementos do BD original.

- SVM realiza a classificação binária.
    - Para estendê-lo para a classificação multiclassse é possível utilizar algumas abordagens (consulte [Han et al., 2012], [Tan et al. 2006])
    - **Exemplo**:
        - Dadas m classes, treinar m classificadores binários (um para cada rótulo classe)
        - Uma tupla de teste será associada ao maior valor positivo entre as previsões de todos os classificadores.

Caso 2: Dados Não-Lineares (1/8)

- O SVM pode ser estendido para o caso em que os dados possuem fronteira de decisão não linear.



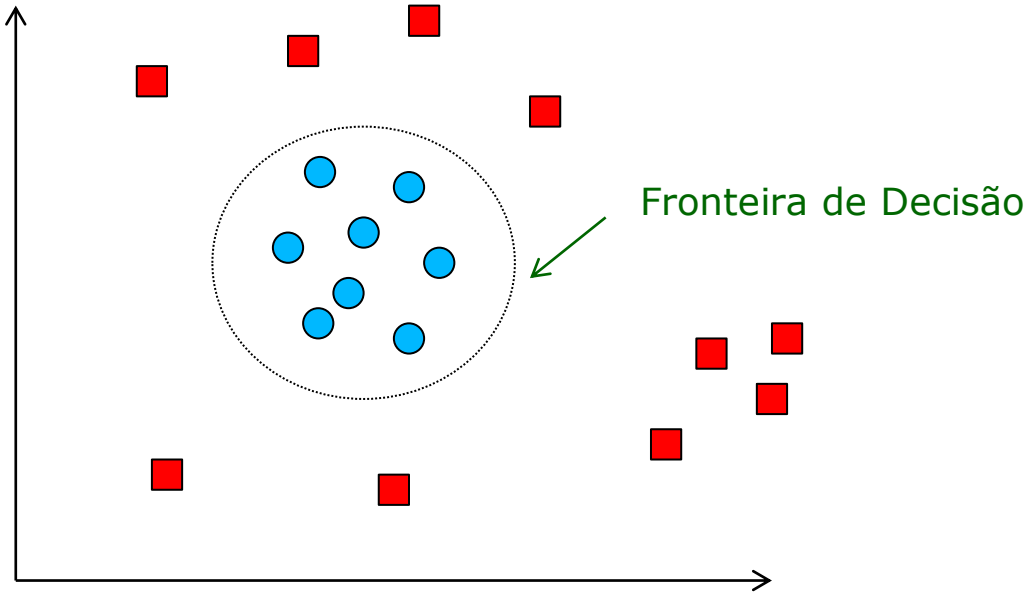
Caso 2: Dados Não-Lineares (2/8)

- Resumidamente, utilizam-se os seguintes passos:
    - **Passo 1**: transformamos os dados de entrada originais para uma dimensão maior utilizando um **mapeamento não linear**.
    - **Passo 2**: procuramos por uma hiperplano separador linear no novo espaço.

- Novamente terminamos com um problema de otimização quadrático que pode ser resolvido usando a formulação linear do SVM.
- A MMH encontrada no novo espaço corresponde a uma hipersuperfície separadora não linear no espaço original.

Caso 2: Dados Não-Lineares (3/8)

- **Transformação Não-Linear**
  - **Exemplo** [Tan et al. 2006]
    - Equação da fronteira: $x^2 - x1 + x1^2 - x2 = -0,46$



    - Uma transformação não linear ϕ é necessária para mapear os dados do espaço original para o novo espaço.
    - Um exemplo seria:
      - $(x1,x2) \rightarrow (x1^2, x^2, 2x1^{1/2}, 2X2^{1/2}, 1)$

Caso 2: Dados Não-Lineares (4/8)

- **Transformação Não-Linear**
  - **Potenciais Problemas**
    - Problema 1: **Como escolher o mapeamento**? Como assegurar que, no novo espaço, haverá uma separação linear?
    - Problema 2: A transformação tem um **custo caro**! Tenho que aplicá-la para todas as instâncias de treino e de teste.
      - Observe que na fórmula para classificar um novo exemplo, existe um produto escalar entre 2 vetores no espaço transformado.

$$d(z) = \sum_{i=1}^{l} yi \times \alpha i \times \Phi(xi) \times \Phi(z) + bo$$

Caso 2: Dados Não-Lineares (5/8)

- **Função Kernel (1/3)**
  - A solução é usar o chamado *kernel trick*. A ideia é a seguinte:
    - Um produto escalar representa a medida de similaridade entre dois vetores.
    - O produto escalar no espaço transformado pode ser expresso como uma função de similaridade no espaço original:

$$K(x,z) = \Phi(x) \times \Phi(z) = (x \times z + 1)$$

Caso 2: Dados Não-Lineares (6/8)

- **Função Kernel (2/3)**
  - K é uma **função kernel**
    - Ela é computada no espaço original.
    - Deve ser uma função que satisfaça o **Teorema de Mercer** [Ng 2012]
      - Toda função que satisfaz esse teorema, garante que haverá uma separação linear no espaço de maior dimensão

$$K(x,z) = \Phi(x) \times \Phi(z) = (x \times z + 1)$$

Caso 2: Dados Não-Lineares (7/8)

- **Função Kernel (3/3)**
  - Alguns exemplos de função Kernel
    - Kernel polinomial de grau p

$$K(x, y) = (x \times y + 1)^p$$

    - Gaussian Radial Basis Function Kernel (RBF)

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

    - Sigmoid Kernel.

$$K(x, y) = \tanh(kx \times y - \delta)$$

Caso 2: Dados Não-Lineares (8/8)

- **Observações**
  - Cada função kernel resulta em um classificador não linear diferente no espaço de entrada original.
  - Segundo [Han et al. 2012] não existem regras de ouro para determinar o melhor kernel (para SVM mais acurado).
    - Na prática, a diferença na acurácia costuma ser pequena.

Comentários Finais

- **Tópicos importantes para quem for trabalhar com o SVM**
  - Resolução do problema de otimização (para achar margem máxima)
  - Teoria do Aprendizado Estatístico

- **Tópicos importantes não mostrados na apresentação**
  - Soft-Margin SVM [Tan et al. 2006]
  - Algoritmo SMO [Ng 2012]

- **Temas de Pesquisa**
  - Melhorar a eficiência das fases de treinamento e teste.
  - Determinar o melhor kernel para um dado dataset
  - Encontrar métodos mais eficientes para o caso multiclasse.

Referências

- **Livros de Mineração de Dados:**
  - J. Han, M. Kamber, J. Pei (2012). "Data Mining: Concepts and Techniques", 3rd Ed.
  - P. N. Tan, M. Stenbach e V. Kumar (2006). "An Introduction to Data Mining"
  - I. W. Witten, E. Frank, M. A. Hall (2011). "Data Mining – Practical Machine Learning Tools and Techniques", 3rd Ed.

- **Material Didático:**
  - Andrew Ng (2012), "CS229 Machine Learning Course Materials", disponível em http://cs229.stanford.edu/

- **Artigos:**
  - C. J. C. Burges (1998), "A Tutorial on Support Vector Machines for Pattern Recognition". Knowledge Discovery and Data Mining 2(2): 1-43.
  - A. Smola e B. Scholkopf (1998). "A Tutorial on Support Vector Regression". Technical Report NC2-TR-1998-030, NeuroCOLT2.