Aprendizagem de Máquina

Máquinas de Vetores de Suporte

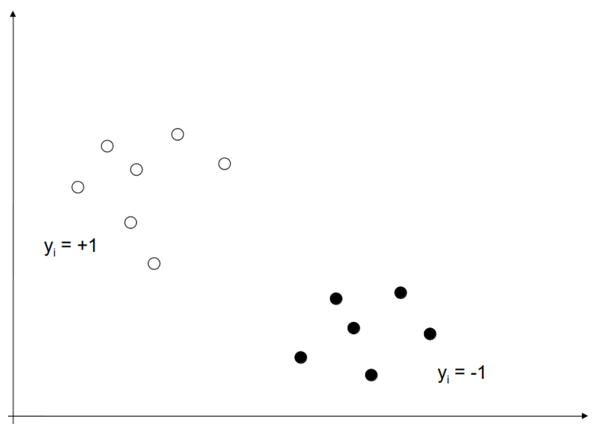
Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB
Departamento de ESTATÍSTICA

Sumário

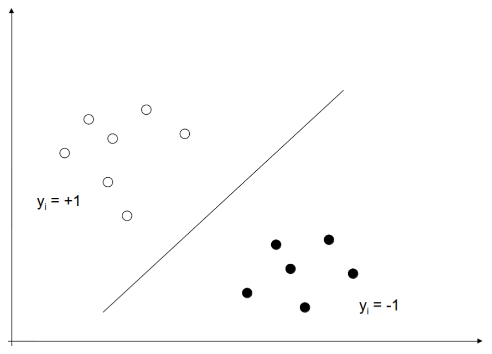
O Hiperplano Ótimo

- Considere um conjunto de N pontos $(i = 1, \ldots, N)$ pertencentes a duas classes $\{+1, -1\}$ linearmente separáveis



O Hiperplano Ótimo

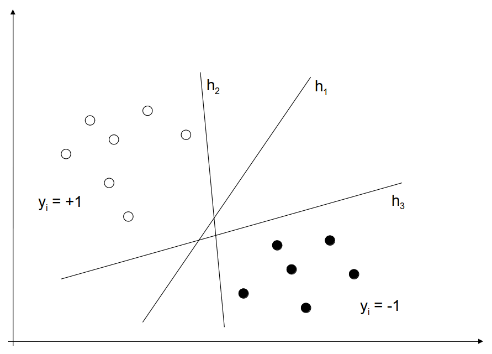
- Como vimos na aula de classificadores lineares, um classificador pode ser construído a partir de um hiperplano de separação $\mathbf{x}\beta + \beta_0 = 0$

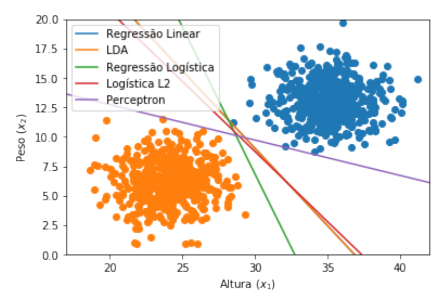


- Se $\mathbf{x}_i \cdot \beta + \beta_0 > 0$
  - $y_i = +1$
- Se $\mathbf{x}_i \cdot \beta + \beta_0 < 0$
  - $y_i = -1$
- Ou $y_i = sign(\mathbf{x}_i \cdot \beta + \beta_0)$

O Hiperplano Ótimo

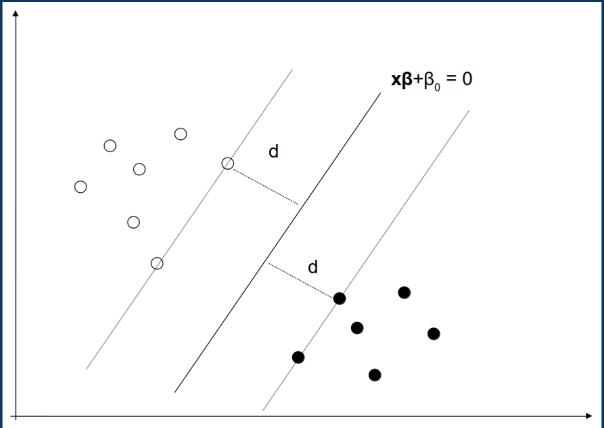
- Existem infinitos hiperplanos que separam dois conjuntos de pontos linearmente separáveis. Assim, qual o melhor?





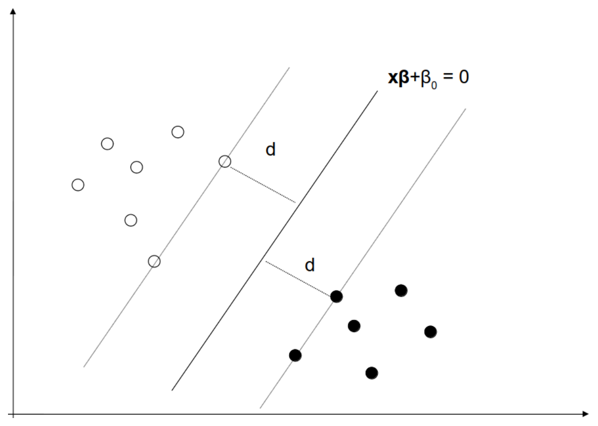
O Hiperplano Ótimo

- O hiperplano ótimo é equidistante às classes e maximiza a **margem** de separação ($2d$)



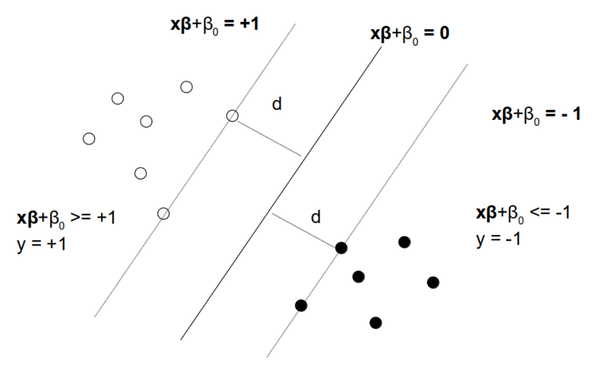
Vetores de Suporte

- Pontos mais próximos do hiperplano ótimo são chamados de **vetores de suporte**



Vetores de Suporte

- Hiperplanos superior e inferior podem ser reescalonados para: $\mathbf{x}_i \cdot \beta + \beta_0 = +1$ e $\mathbf{x}_i \cdot \beta + \beta_0 = -1$
- Margem $2d$ é calculada como: $\frac{2}{||\beta||}$



Problema

- Maximizar a margem

$$\frac{2}{||\beta||}$$

- Sujeito a:

$$\mathbf{x}_i \cdot \beta + \beta_0 \geq +1, \text{se } y_i = +1$$
$$\mathbf{x}_i \cdot \beta + \beta_0 \leq -1, \text{se } y_i = -1$$

- Ou minimizar

$$\frac{||\beta||^2}{2}$$

- Sujeito a:

$$y_i(\mathbf{x}_i \cdot \beta + \beta_0) - 1 \geq 0$$

Multiplicadores de Lagrange

- Minimizar
$$\frac{||\beta||^2}{2}$$
- Sujeito a:
$$y_i(\mathbf{x}_i \cdot \beta + \beta_0) - 1 \geq 0$$
- Por Multiplicadores de Lagrange, equivale a minimizar:
$$L = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathbf{x}_i \cdot \beta + \beta_0) - 1 \right]$$
- Multiplicador de Lagrange $\alpha_i \geq 0$ pode ser visto como a “força” da i-ésima restrição

Multiplicadores de Lagrange

- Algoritmo Sequential Minimal Optimization (SMO – LibSVM) retorna solução única e ótima para os valores de α_i
- Existe um α_i para cada exemplo de treinamento
- Na solução ótima, $\alpha_i > 0$ para os vetores suporte e $\alpha_i = 0$ para os outros exemplos !!!
- Intuição: O hiperplano ótimo depende apenas dos vetores suporte

SVM com Margem Suave

SVM com Margem Suave

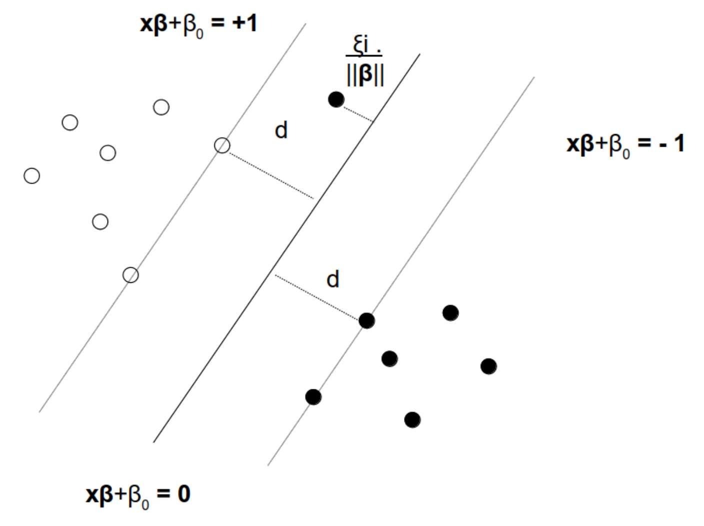
- A formulação anterior é definida para conjuntos perfeitamente linearmente separáveis (SVM de margem rígida)
- Para conjuntos não-linearmente separáveis pequenos erros podem ser tolerados
-
- Minimizar

$$\frac{||\beta||^2}{2} + C\sum_{i=1}^{N} \xi_i$$

- Sujeito a:

$$y_i(\mathbf{x}_i \cdot \beta + \beta_0) - 1 + \xi_i \geq 0$$

SVM com Margem Suave

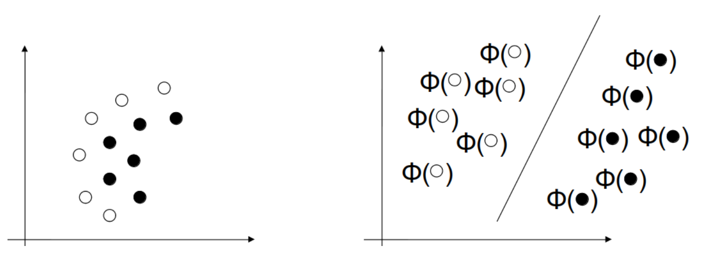


SVM não-Linear

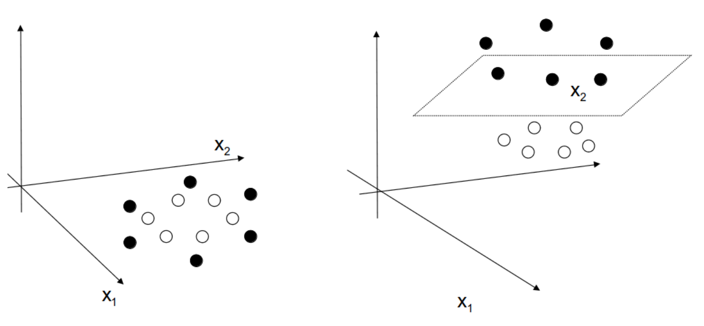
SVM não-Linear

- SVM **linear** ainda é limitado mesmo com margens flexíveis
- Podemos fazer uma generalização não-linear de SVM
  - Mapear espaço original para espaço não-linear de **maior dimensão** onde exemplos sejam linearmente separáveis
  - Construir **hiperplano ótimo** no novo espaço

SVM não-Linear



SVM não-Linear



SVM (Kernels)

- Em SVMs não-lineares, pontos são mapeamentos implicitamente através de uma **função de Kernel**
  - Kernel polinomial (parâmetro p): $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j + 1)^p$
  - Kernel RBF (parâmetro γ): $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2}$
- A escolha do Kernel é importante para o desempenho das SVMs
- Dependendo do Kernel utilizado alguns parâmetros devem ser definidos
- Kernel RBF é mais flexível que o polinomial
- Kernel RBF depende de parâmetro γ
- Valores altos dão maior flexibilidade ao modelo mas também aumentam risco de overfitting

SVM (Kernels)

- Sobre parâmetro C:
  - Valores muito altos propiciam a geração de modelos mais complexos (risco de overfitting)
  - Valores muito baixos podem aumentar risco de underfitting

SVM para Regressão

SVM para Regressão (SVR)

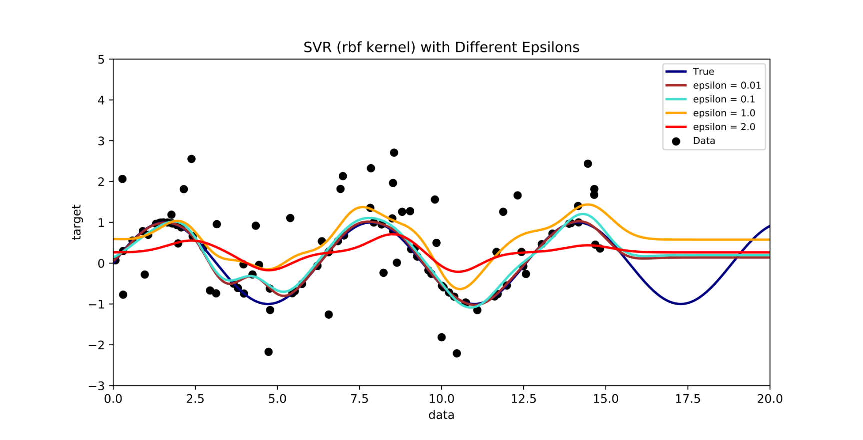
- O modelo produzido pelo SVM para classificação depende só do conjunto de vetores de suporte e ignora as instâncias que ficam além da margem
- De forma similar, o SVR depende apenas de um subconjunto dos dados de treinamento, ignorando as instâncias cuja estimativa de y é bem aproximada
- Minimizar

$$\frac{||\beta||^2}{2}$$

- Sujeito a:

$$|y_i - (\mathbf{x}_i \cdot \beta + \beta_0)| \leq \epsilon$$

SVM para Regressão



Para Terminar

- SVM se situa dentre as técnicas de aprendizado mais poderosas
- Baseada em uma teoria forte
- Ou seja, justificável teoricamente e com bom desempenho empírico
- Apesar de ter poucos parâmetros para selecionar (e.g., função de kernel), escolha adequada é importante
- Maior desvantagem é o tempo de treinamento e uso
- SVM não estima probabilidades de classe
  - Mais sobre isso mais à frente
- O truque do Kernel pode ser usado em outros algoritmos

Sugestão de Atividade

- Baixe os scripts (Python e R) disponibilizados no SIGAA para geração de conjuntos artificiais
- Gere conjuntos e brinque com Kernels/parâmetros diferentes e veja o quanto eles se aproximam de classificar/estimar corretamente os dados (use gráficos de dispersão com cores diferentes para as classes)
- Exemplo de uso:
  - gera_conjunto_classificacao('teste', 1000, 2, 'pol1', 0.1)
  - gera_conjunto_regressao('teste', 1000, 2, 'sin3', 0.1)

FIM

Aprendizagem de Máquina

Máquinas de Vetores de Suporte

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA