Aprendizado de Máquina

Support Vector Machines

Luiz Eduardo S. Oliveira

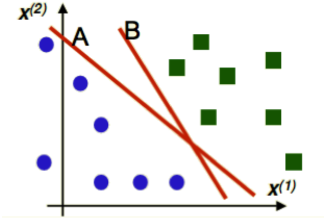
Universidade Federal do Paraná
Departamento de Informática
web.inf.ufpr.br/luizoliveira

Introdução

- Classificador linear proposto por Vapnik e Chervonenkis em 1963
- Vapnik et al, 1995 (Kernel trick and Soft Margin)
- Utilizado com sucesso para resolver diferentes problemas de classificação e regressão.

Introdução

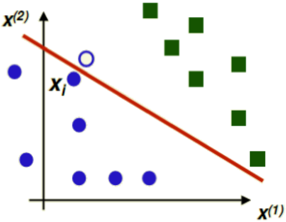
- Como vimos anteriormente, existem diferentes fronteiras que separam dados linearmente separáveis.
- Perceptron é capaz de encontrar essas fronteiras.



- Qual fronteira deve ser escolhida?

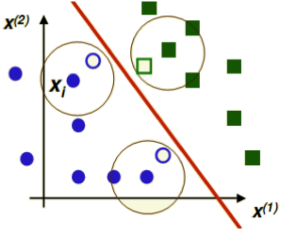
Introdução

- Suponha que a fronteira escolhida é a A
- Como ela está bem próxima da classe azul, **seu poder de generalização é baixo**
- Note que um novo elemento (dados não usados no treinamento), bem próximo de um azul será classificado erroneamente



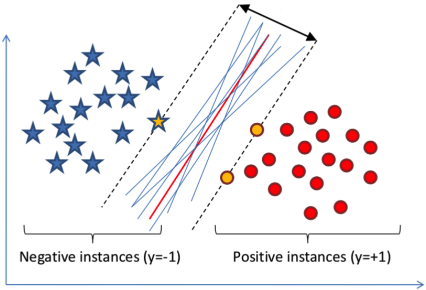
Introdução

- Escolhendo a fronteira B, podemos notar que o pode de generalização é bem melhor.
- Novos dados são corretamente classificados, pois temos uma fronteira mais distante dos dados de treinamento



SVM - Hard Margin

- Dado um conjunto de treinamento $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_n \in R^n$, $y_1, y_2, \ldots, y_n \in \{+1, = 1\}$
- Encontrar um hyperplano que separe as instâncias positivas das negativas, maximizando a margem entre essas duas classes de instâncias.
- Os pontos que definem essa margem máxima são conhecidos como vetores de suporte.



SVM - Hard Margin

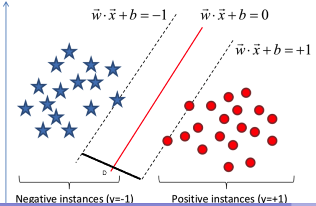
- A margem é a distância entre os dois hiperplanos paralelos:

$$\vec{w} \cdot \vec{x} + b = -1 \text{ e } \vec{w} \cdot \vec{x} + b = +1$$

- ou equivalente,

$$\vec{w} \cdot \vec{x} + (b+1) = 0 \text{ e } \vec{w} \cdot \vec{x} + (b-1) = 0$$

- A distância (D) entre as margens é dada por $|b_1 - b_2|/||\vec{v}||$, ou seja, $\frac{2}{||\vec{v}||}$
- A distância de qualquer margem para o hiperplano é $\frac{1}{||\vec{v}||}$
- Desta forma, para maximizar a margem precisamos minimizar $||\vec{v}||$, ou equivalente $\frac{1}{2}||\vec{v}||^2$



SVM - Hard Margin

- É necessário ainda adicionar as restrições para que todas as instâncias sejam corretamente classificadas:

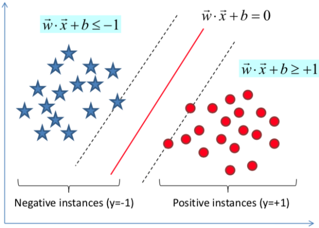
$$\vec{w} \cdot \vec{x} + b \leqslant -1 \text{ se } y_i = -1$$

$$\vec{w} \cdot \vec{x} + b \geqslant +1 \text{ se } y_i = +1$$

- ou equivalente,

$$y_i(\vec{w} \cdot \vec{x} + b) \geqslant 1$$

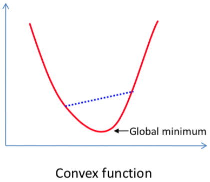
- Em resumo, devemos minimizar $\frac{1}{2}||\vec{v}||^2$ sujeito à $y_i(\vec{w} \cdot \vec{x} + b) \geqslant 1$, para $i = 1, \ldots, N$

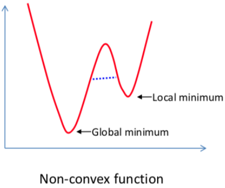


SVM - Problema de Otimização

$$\text{Minimizar } \frac{1}{2}\sum_{i=1}^{n} w_i^2 \text{ sujeito à } y_i(\vec{w}\cdot\vec{x}+b) \geqslant 1, \text{ para } i = 1, \ldots, N$$

- Chamado de **forma primal** para SVM lineares
- Problema convexo de otimização quadrática com n variáveis ($w_i, i = 1, \ldots, n$) em que n é o número de características na base de dados.
- Podem ser resolvidos de forma eficiente pois um mínimo local é também um mínimo global.





SVM - Problema de Otimização

- Problemas desse tipo podem ser solucionados com a introdução de uma função Lagrangiana, que engloba as restrições à função objetivo, associadas a parâmetro denominados multiplicadores de Lagrange α_i

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\sum_{i=1}^{n} w_i^2 - \sum_{i=1}^{n} \alpha_i (y_i(\vec{w} \cdot \vec{x_i} + b) - 1)$$

- Isso nos leva a **forma dual**, que continua um problema de otimização quadrática mas com N variáveis ($\alpha_i, i = 1 \ldots, N$) em que N é o número de exemplos da base (e não o número de características)

$$\text{Maximizar} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\vec{x_i} \cdot \vec{x_j}) \text{ sujeito à } \alpha_i \geqslant 0 \text{ e } \sum_{i=1}^{N} \alpha_i y_i = 0$$

SVM - Problema de Otimização

Vantagens da forma dual

- Permite a representação do problema de otimização em termos de produtos internos entre os dados (útil na posterior não linearização do SVM)
- Número de parâmetros é limitado pelo número de vetores de suporte e não pela quantidade de características.
- Útil em problemas de grandes dimensões.
- O classificador SVM é dado por

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i \vec{x}_i \cdot \vec{x} + b\right)$$

Esta função linear representa o hiperplano que separa os dados com maior margem.

Exemplo

- Considere o SVM abaixo em que bias encontrado para o hiperplano é -1.667. Classifique a instância $[3,3]^T$ e atribua o exemplo a classe positiva ou negativa.

i	x_i	y_i	α_i
1	$[1,3]^T$	-1	0
2	$[2,1]^T$	-1	0.33
3	$[4,5]^T$	-1	0
4	$[6,7]^T$	-1	0
5	$[8,7]^T$	-1	0.11
6	$[5,1]^T$	1	0.44
7	$[7,1]^T$	1	0
8	$[9,4]^T$	1	0
9	$[12,7]^T$	1	0
10	$[13,6]^T$	1	0

0.33 * (-1) * (2*3+1*3) + 0.11 * (-1) * (8*3+7*3) + 0.44*1*(5*3+1*3)+ -1.667 = -1.667

SVM - Soft Margin

- E quando os dados não são linearmente separáveis?
- Por exemplo, presença de outliers, erros na extração de características, etc.

SVM - Soft Margin

Solução: Variáveis de folga

- Permitir que alguns dados possam violar a restrição imposta no problema de otimização.
- Atribuir uma variável de folga (slack variable) ξ_i. Essas variáveis relaxam as restrições impostas ao problema de otimização.
- O termo C é utilizado para decidir o quanto penalizar os pontos classificados erroneamente.

Primal

$$\text{Minimizar } \frac{1}{2}\sum_{i=1}^{n} w_i^2 + C\sum_{i=1}^{N}\xi_i \text{ sujeito à } y_i(\vec{w}\cdot\vec{x}+b) \geqslant 1-\xi_i, \text{ para } i = 1,\ldots,N$$

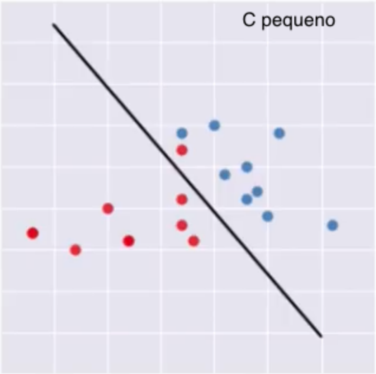
Dual

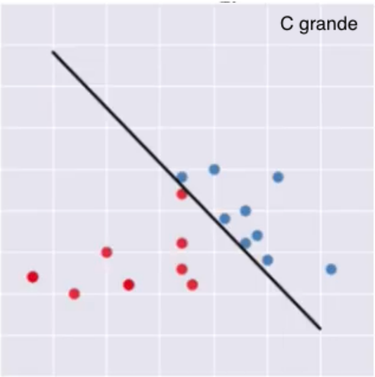
$$\text{Maximizar} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j(\vec{x_i}\cdot\vec{x_j}) \text{ sujeito à } 0 \leqslant \alpha_i \leqslant C \text{ e } \sum_{i=1}^{N}\alpha_i y_i = 0$$

SVM - Soft Margin

Impacto do parâmetro C

- C pequeno: priorizar a simplicidade (soft margin)
  - C muito pequeno pode ocasionar under-fitting
- C grande: priorizar errar pouco no treinamento.
  - Maior risco de over-fitting.



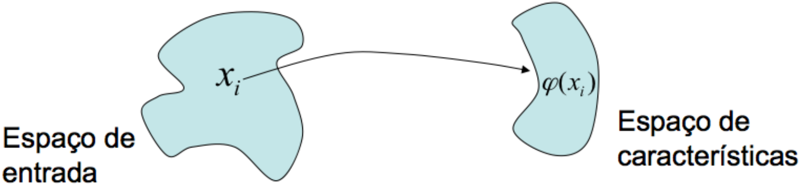


SVM - Mapeamento não linear

- A grande maioria dos problemas reais não são linearmente separáveis.
- A pergunta então é: "Como resolver problemas que não são linearmente separáveis com um classificador linear?"

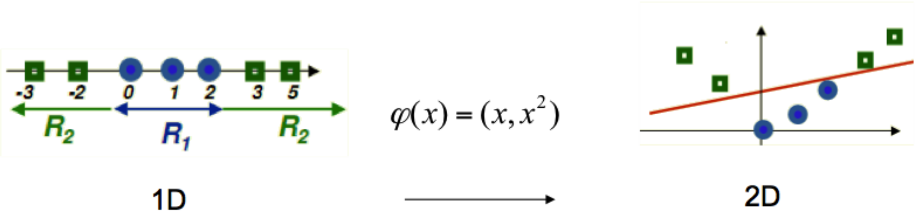
Kernel Trick

Projetar os dados em um espaço onde os dados são linearmente separáveis



SVM - Mapeamento não linear

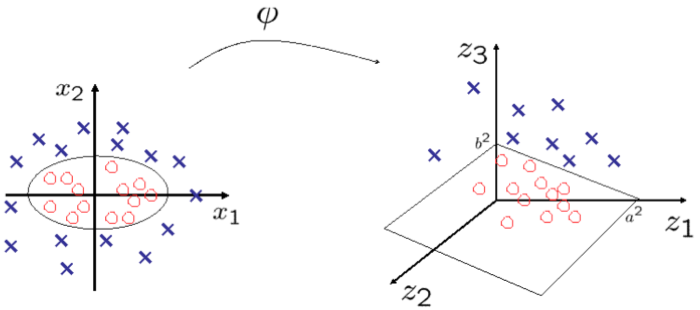
- A função que projeta o espaço de entrada no espaço de características é conhecida como Kernel
- Encontrar um hiperplano que separe os dados nesse espaço.
- Por exemplo:



SVM - Mapeamento não linear

Outro exemplo

$$\begin{array}{rccl} \phi : & \Re^2 & \longrightarrow & \Re^3 \\ & (x_1, x_2) & \longmapsto & (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{array}$$



SVM - Mapeamento não linear

$$f(\vec{x}) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(\vec{x_i} \cdot \vec{x}) + b\right)$$

- K é a função de Kernel a qual precisa satisfazer algumas condições (Teorema de Mercer)
  - Poder ser resolvido pela otimização quadrática.
- Um kernel é o produto interno em algum espaço de caracterísitcas

$$K(\vec{x_i} \cdot \vec{x_j}) = \Phi(\vec{x_i}) \cdot \Phi(\vec{x_j})$$

Alguns exemplos de kernel

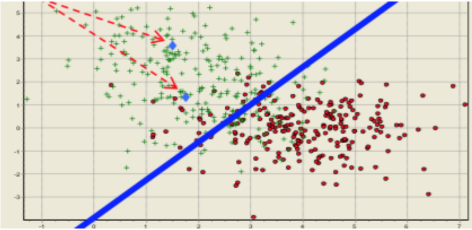
$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j$	Linear kernel
$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \lVert \vec{x}_i - \vec{x}_j \rVert^2)$	Gaussian kernel
$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q$	Polynomial kernel
$K(\vec{x}_i, \vec{x}_j) = \tanh(k\vec{x}_i \cdot \vec{x}_j - \delta)$	Sigmoidal

Escolhendo um Kernel

- Kernel Linear
  - Problemas linearmente separáveis
  - Vetores de características esparsos de altas dimensões
  - Treinamento mais rápido com menos parâmetros.
- Kernel RBF
  - Problemas que não são linearmente separáveis
  - Em geral produz resultados melhores do que os outros kernels não lineares.

Estimando Probabilidades

- A função $f(x)$ permite atribuir o exemplo a uma determinada classe.
- Entretanto, alguns pontos estão mais próximos da fronteira do que outros.
- Em alguns casos, é interessante uma probabilidade associada a classificação de um dado padrão.

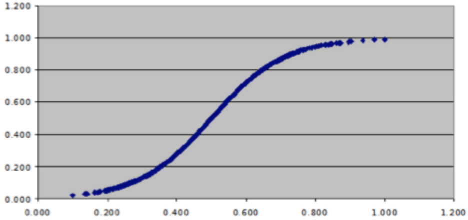


Estimando Probabilidades

Método de Platt (Platt Scaling)

$$P(y = 1|x) = \frac{1}{1 + \exp(Af(f) + B)}$$

- Transformação logística da saída do classificador f(x) em que A e B são dois escalares estimados na base de treinamento (e.g., regressão logística).



SVM Multi-classe

- SVM são classificadores binários, ou seja, separam duas classes.
- Entretanto, a grande maioria dos problemas reais possuem mais que duas classes.
- Como utilizar o SVM nesses casos?
  - Pairwise e Um-Contra-Todos

SVM Multi-classe

Pairwise

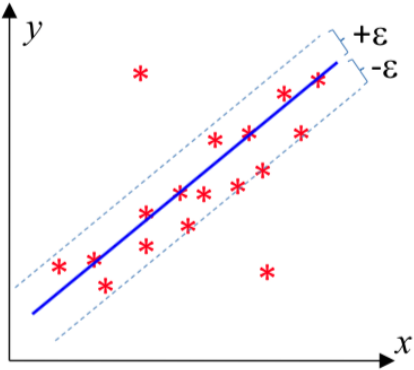
- Construir $k(k-1)/2$ classificadores para discriminar cada par de classes, obtendo assim $k(k-1)/2$ classificadores $f_{k,j}(x)$
- Escolher a classe escolhida pela maioria dos "pairwise" SVMs.
- Para 10 classes, são necessários 45 SVMs

Um-Contra-Todos

- k classificadores que discriminam uma classe contra todas as outras.
- Escolher a classe com o maior score, ou seja $y = \arg\max_k f_k(x)$

SVR (Regressão)

- Encontrar uma função $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$ que aproxime $y_1, \ldots, y_N$
- Ter no máximo ϵ desvio dos valores de y_i
- Ser o mais "reto" possível para evitar overfitting.

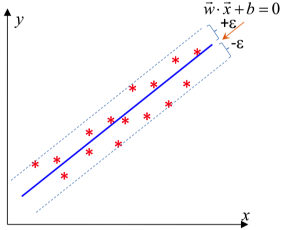


SVR (Regressão)

Formulação hard-margin SVR

- Encontrar $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$ minimizando $\frac{1}{2}||\vec{w}||^2$ sujeito à

$$y_i - (\vec{w} \cdot \vec{x} + b) \leqslant \epsilon \text{ e } y_i - (\vec{w} \cdot \vec{x} + b) \geqslant -\epsilon \text{ para } i = 1, \ldots, N$$

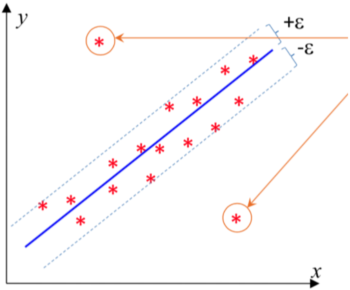


A diferença entre y_i e a função encontrada deve ser menor que ϵ e maior que $-\epsilon$. Ou seja, todos os pontos y_i deve estar dentro da "ϵ-fronteira".

SVR (Regressão)

Formulação soft-margin SVR

- Se tivermos pontos como esses (outliers ou ruído) podemos:
  - Aumentar ϵ para garantir que esses pontos fiquem dentro da "ϵ-fronteira"
  - Atribuir variáveis de folga para cada ponto, da mesma maneira que fizemos no "soft-margin" SVM.

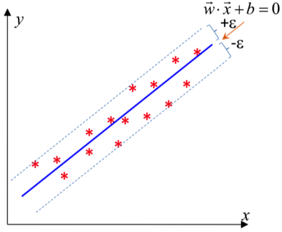


SVR (Regressão)

Formulação soft-margin SVR

- Encontrar $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$ minimizando $\frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$ sujeito à

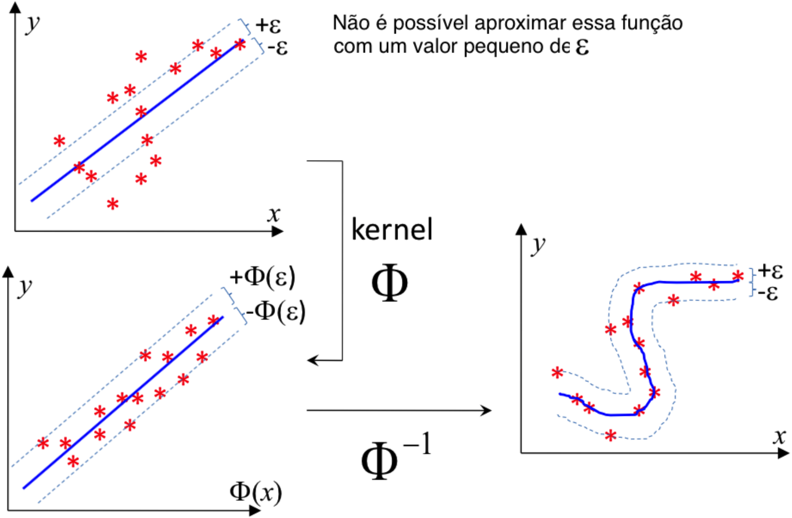
$$y_i - (\vec{w} \cdot \vec{x} + b) \leqslant \epsilon + \xi_i \text{ e } y_i - (\vec{w} \cdot \vec{x} + b) \geqslant -\epsilon - \xi_i^* \text{ para } i = 1, \ldots, N$$



Pontos fora da "ϵ-fronteira" são penalizados.

SVR (Regressão)

SVR Não linear



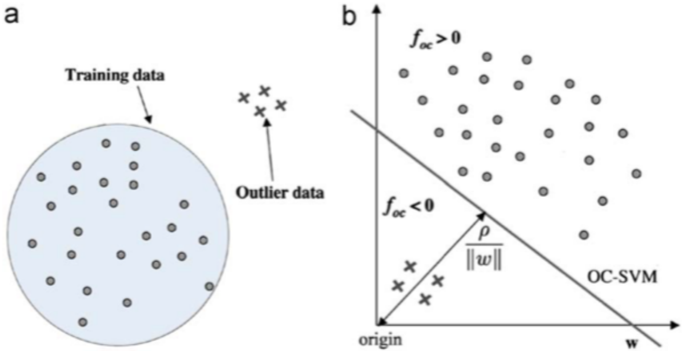
Não é possível aproximar essa função com um valor pequeno de ε

One-Class SVM

- Suponha a seguinte situação
  - Queremos determinar se um determinado padrão pertence a uma determinada classe ou não
  - Por exemplo, um sistema de monitoramento no qual a tarefa é determinar se algo anormal aconteceu.
  - Detectar uma anomalia
- Nesses casos temos muitos dados de treinamento de uma única classe.
- A anomalia não ocorre com muita frequência (como diz o próprio nome)
- Desafio
  - Construir um classificador com uma única classe.

One-Class SVM

- One-Class SVM basicamente separa todos os pontos da origem e maximiza a distância do hiperplano da origem.
- Desta forma, o rótulo +1 é atribuído aos exemplo que estão contidos dentro da pequena região dos exemplos conhecidos.
- -1 é o rótulo atribuído para tudo que está fora da região modelada.



One-Class SVM

SVDD - Support Vector Data Description

- Encontrar a menor hiper-esfera que contem todos os exemplos da classe modelada.