SVM

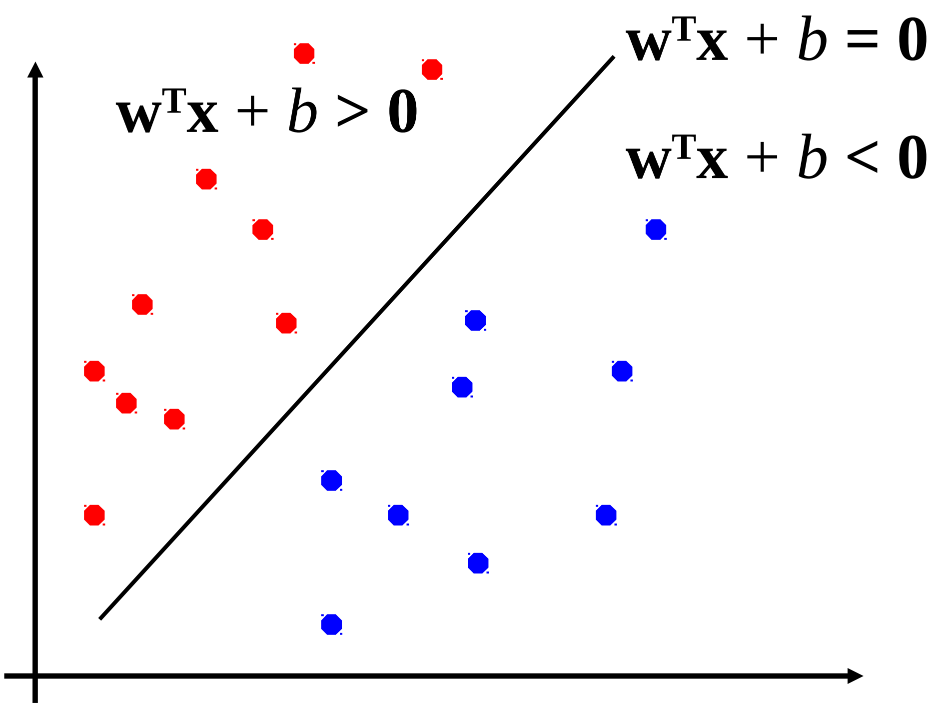
Máquinas de Vetores de Suporte

Prof. Frederico Coelho.

(Baseado no livro de Antônio de Pádua Braga e
No material da Universidade do texas em Austin)

Introdução

Classificação > tarefa de separar classes no espaço de entrada



$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$$

Introdução

Dado um conjunto de dados $D_L = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$;

Onde $y_i \in \{-1,+1\}$

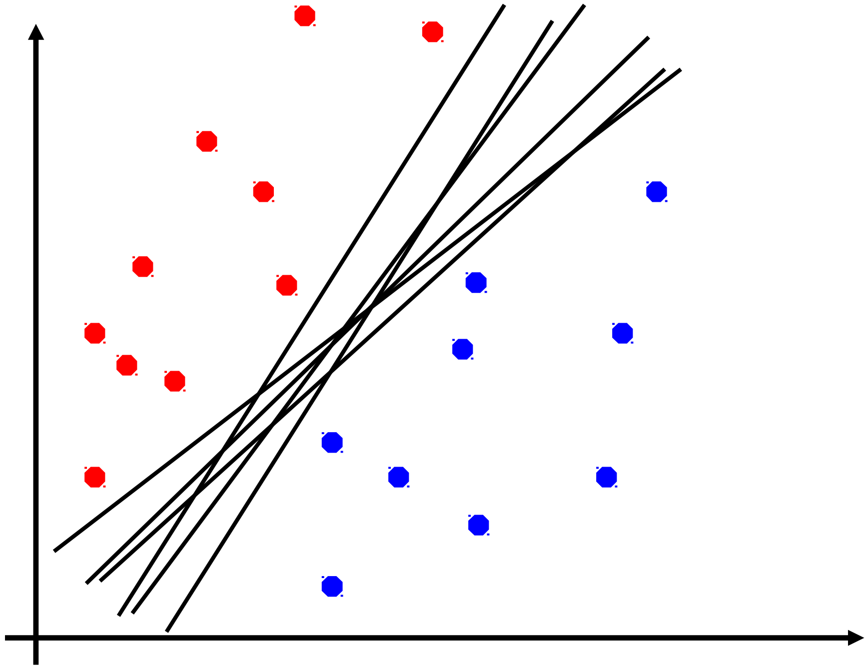
Se $\text{sign}(\mathbf{w}^T\mathbf{x} + b)$ tiver o mesmo sinal de y então a classificação estará correta

Assim, para que todos os vetores x_i sejam classificados corretamente a seguinte desigualdade deve ser satisfeita para todos os N pares (x_i,y_i)

$$\mathbf{y}(\mathbf{w}^T\mathbf{x} + b) \geq 1$$

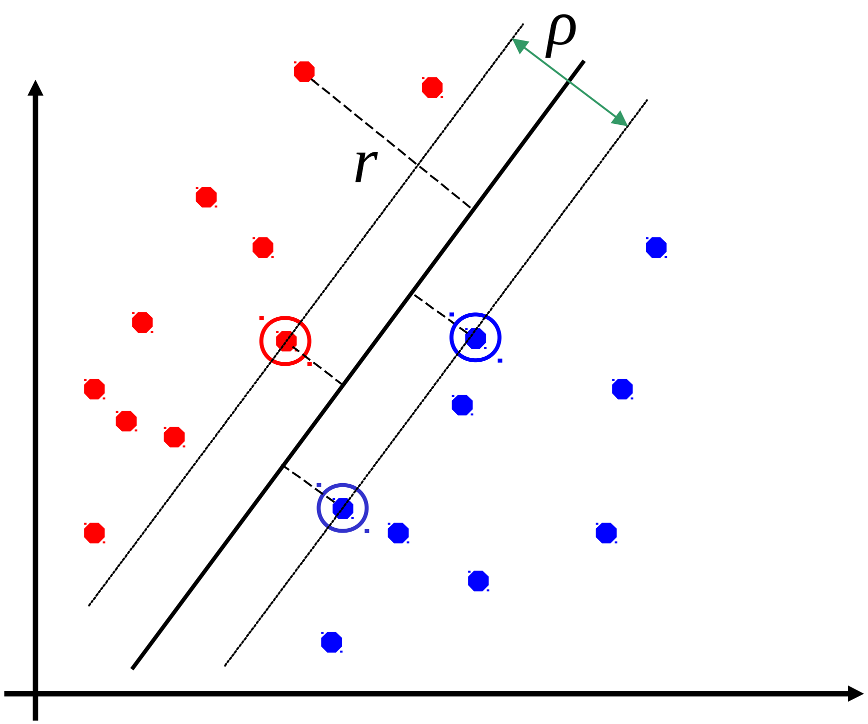
Separador linear

Mas qual separador é o melhor?



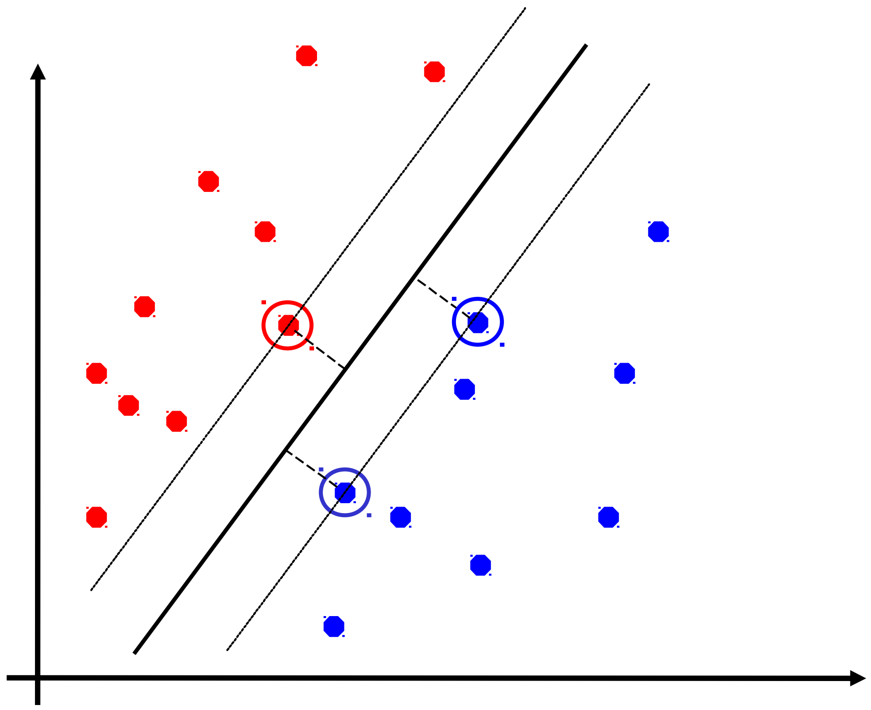
Margem de classificação

- A distância de uma amostra x_i ao separador é $r = \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|}$
- Amostras próximas ao hiperplano são chamados Vetores de Suporte
- ρ é a margem máxima entre vetores de suporte.



Margem larga

Apenas vetores de suporte interessam – todo o resto pode ser ignorado



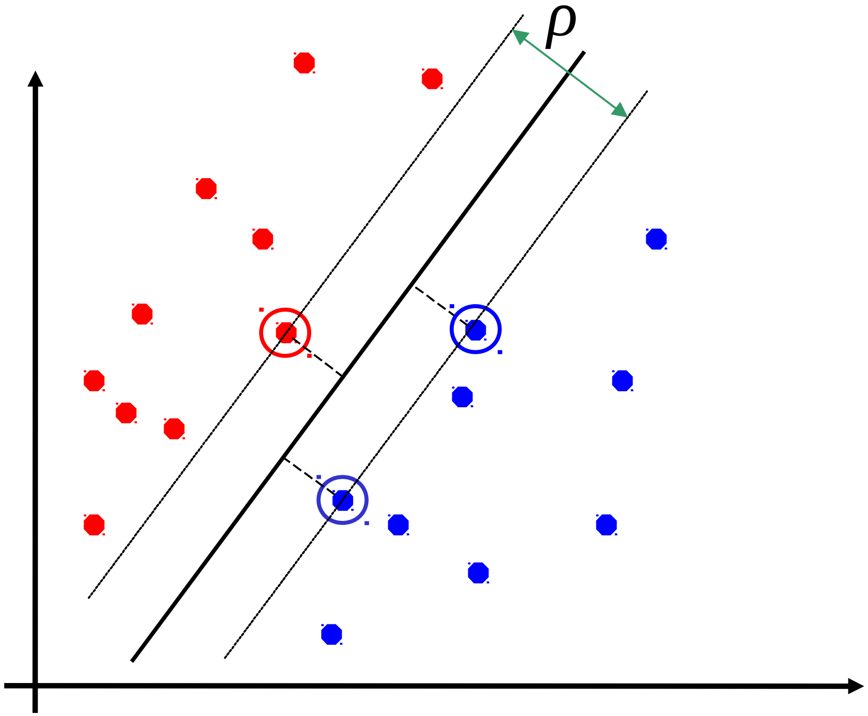
Objetivo – maximizar a margem

SVM linear

Dado um conjunto de dados $D_L = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$:

Onde $y_i \in \{-1,+1\}$, para cada par (x_i, y_i) teremos:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b &\le -\rho/2 \quad \text{se } y_i = -1 \\ \mathbf{w}^T\mathbf{x}_i + b &\ge \rho/2 \quad \text{se } y_i = 1 \end{aligned} \quad \Leftrightarrow \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge \rho/2$$



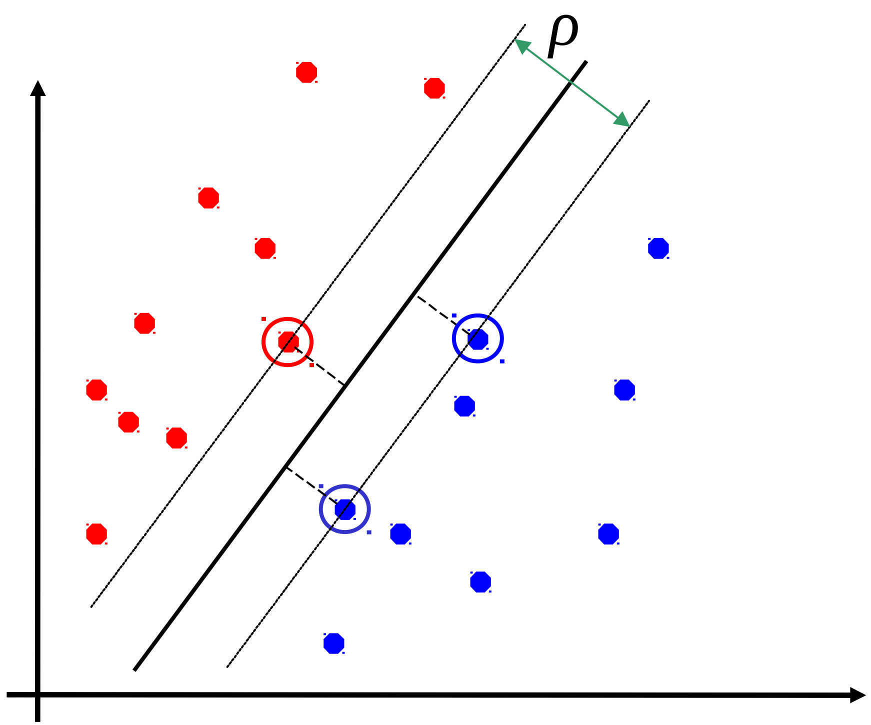
SVM linear

Para cada vetor de suporte Xs a desigualdade anterior é uma igualdade.

Reescalando w e b por ρ/2 temos que a distância entre o hiperplano e Xs será:

$$r = \frac{y_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

Então a margem será: $\rho = 2r = \frac{2}{\|\mathbf{w}\|}$



SVM linear

Então podemos formular o seguinte problema quadrático de otimização:

Encontrar w e b tal que

$\rho = \frac{2}{\|\mathbf{w}\|}$ é maximizado,

Para todo (x_i,y_i), i=1,...,n : $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Que por sua vez pode ser reformulado como:

Encontrar w e b tal que

$\Phi(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$ é minimizado,

Para todo (x_i,y_i), i=1,...,n : $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Resolvendo o problema de otimização

Encontrar w e b tal que

$\mathbf{\Phi}(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$, é minimizado,

Para todo (x_i, y_i), i=1,...,n : $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

- Necessidade de otimizar uma função quadrática sujeita a restrições lineares.
- Problemas de otimização quadrática são uma classe bem conhecida de problemas de programação matemática para os quais existem vários algoritmos (não triviais).
- A solução envolve a construção de um problema duplo, onde um multiplicador de Lagrange αi está associado a todas as restrições de desigualdade no problema primal (original):

Encontrar $\alpha_1 \ldots \alpha_n$ tal que

$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - ½\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$ é máximo e

- $\Sigma\alpha_i y_i = 0$
- $\alpha_i \geq 0$ para todo α_i

Resolvendo o problema de otimização

- Dada uma solução $\alpha_1 \ldots \alpha_n$ para o problema duplo, a solução para o primal é:

$$\boxed{\mathbf{w} = \Sigma \alpha_i y_i \mathbf{x}_i \qquad b = y_j - \Sigma \alpha_i y_i \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \quad \text{para todo } \alpha_j > 0}$$

- Cada α_i diferente de zero indica que x_i correspondente é um vetor de suporte.

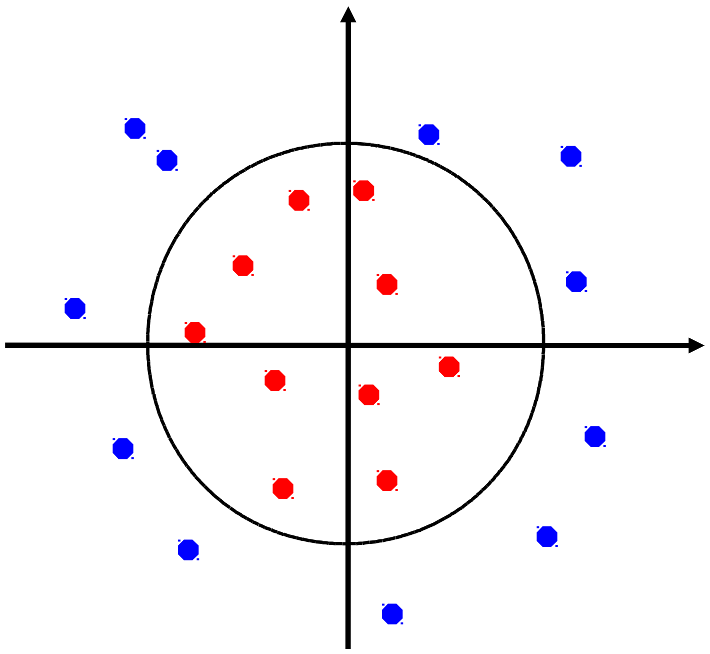
Em seguida, a função de classificação é (observe que não precisamos de w explicitamente):

$$\boxed{\hat{y} = \Sigma \alpha_i y_i \mathbf{x}_i^{\mathrm{T}} \mathbf{x} + b}$$

- Observe que ele depende de um produto interno entre o ponto de teste x e os vetores de suporte x_i .Lembre-se também de que a solução do problema de otimização envolveu a computação dos produtos internos $x_i^T x_j$ entre todos os pontos de treinamento.

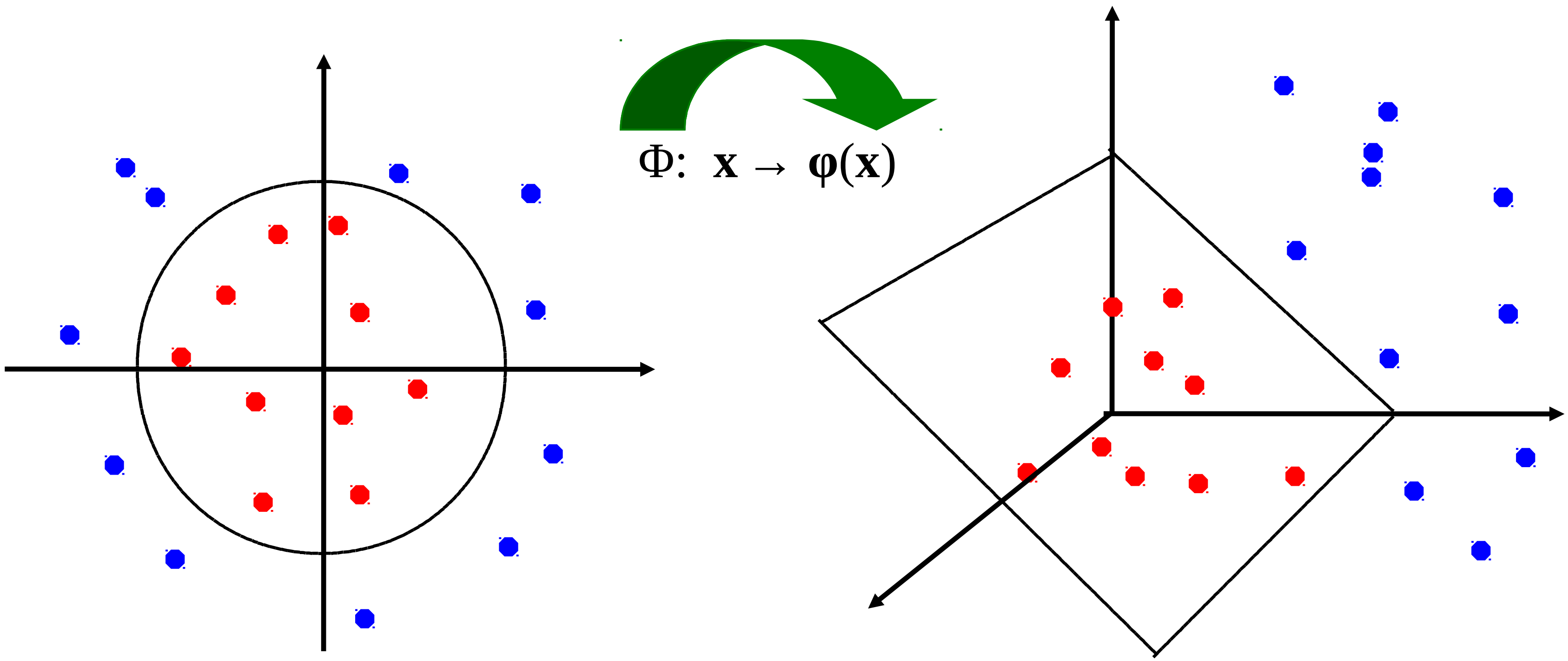
SVM não-linear

E se o problema não for linearmente separável?



SVM não-linear

As amostras X no espaço de entrada podem ser mapeadas para outro espaço onde o problema pode ser linearmente separável.



SVM não-linear

O classificador linear depende do produto interno entre os vetores
$K(x_i, x_j) = x_i^T x_j$

Se todo ponto de dados for mapeado no espaço de alta dimensão através de alguma transformação $\Phi: x \rightarrow \varphi(x)$, o produto interno se tornará:
$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$

Uma função do kernel é uma função que é equivalente a um produto interno em algum espaço de variáveis

Assim, uma função do kernel mapeia implicitamente os dados para um espaço de alta dimensão (sem a necessidade de calcular cada $\varphi(x)$ explicitamente).

SVM não-linear

Para algumas funções K (x_i, x_j), verifcar se K $(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ pode ser complicado.

Teorema de Mercer:
Toda função simétrica definida semi-positiva é um kernel

As funções simétricas definidas semi-positivas correspondem a uma matriz simétrica definida semi-positiva:

$$
K = \begin{array}{|c|c|c|c|c|}
\hline
K(\mathbf{x}_1,\mathbf{x}_1) & K(\mathbf{x}_1,\mathbf{x}_2) & K(\mathbf{x}_1,\mathbf{x}_3) & \cdots & K(\mathbf{x}_1,\mathbf{x}_n) \\
\hline
K(\mathbf{x}_2,\mathbf{x}_1) & K(\mathbf{x}_2,\mathbf{x}_2) & K(\mathbf{x}_2,\mathbf{x}_3) & & K(\mathbf{x}_2,\mathbf{x}_n) \\
\hline
 & & & & \\
\hline
\cdots & \cdots & \cdots & \cdots & \cdots \\
\hline
K(\mathbf{x}_n,\mathbf{x}_1) & K(\mathbf{x}_n,\mathbf{x}_2) & K(\mathbf{x}_n,\mathbf{x}_3) & \cdots & K(\mathbf{x}_n,\mathbf{x}_n) \\
\hline
\end{array}
$$

SVM não-linear

Linear: $K(x_i, x_j) = x_i^T x_j$

Mapeamento Φ: $x \to \varphi(x)$, onde $\varphi(x)$ é o próprio x

Polinômio de potência p: $K(x_i, x_j) = (1 + x_i^T x_j)^p$

Mapeamento Φ: $x \to \varphi(x)$, onde $\varphi(x)$ tem $\binom{d+p}{p}$ dimensões

Gaussiana (função de base radial): $K(x_i, x_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$

Mapeamento Φ: $x \to \varphi(x)$, onde $\varphi(x)$ é de dimensão infinita: todo ponto é mapeado para uma função (um gaussiano); A combinação de funções para vetores de suporte é o separador.

SVM não-linear

Então o problema dual pode ser formulado como:

Encontrar $\alpha_1 \ldots \alpha_n$ tal que

$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ é máximo e

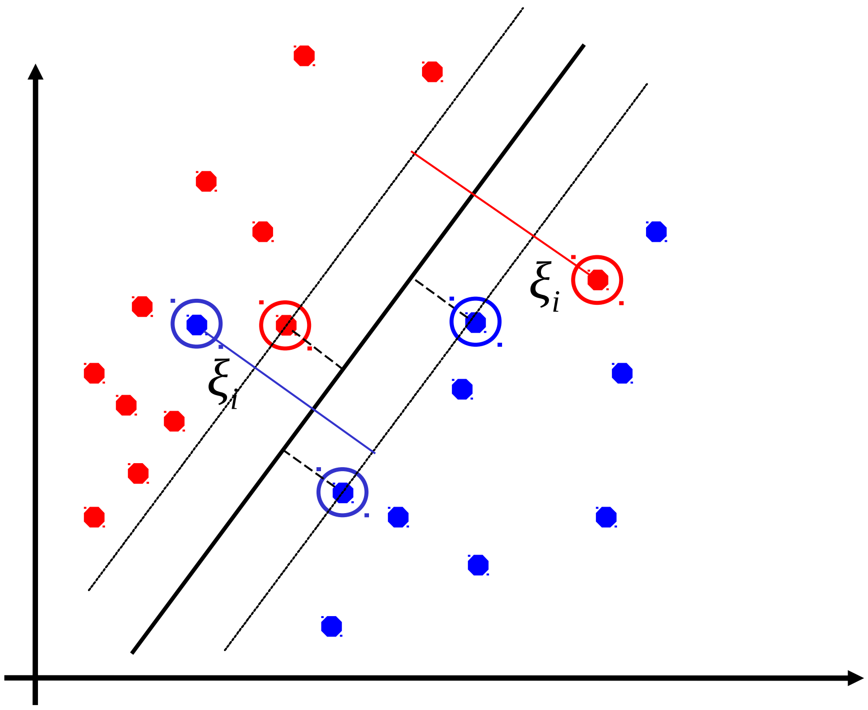
- $\Sigma\alpha_i y_i = 0$
- $\alpha_i \geq 0$ para todo α_i

E a solução será:

$$\boxed{f(\mathbf{x}) = \Sigma\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b}$$

SVM com variável de folga

Variáveis de folga ξi podem ser adicionadas para permitir a classificação incorreta de exemplos difíceis ou ruidosos, a margem resultante é denominada suave (soft margin).



SVM linear com variável de folga

O problema pode ser formulado da seguinte forma:

Encontrar **w** e b tal que

$\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{w} + C\Sigma\xi_i$ é mínimo

e para todo $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$

O parâmetro C pode ser visto como uma maneira de controlar o *overfitting*: "trade off" entre a importância de maximizar a margem e ajustar os dados de treinamento.

SVM linear com variável de folga

Resolvendo o problema dual teremos:

$$\boxed{\begin{array}{l} \mathbf{w} = \Sigma\alpha_i y_i \mathbf{x}_i \\ b = y_k(1 - \xi_k) - \Sigma\alpha_i y_i \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k \quad \text{para qualquer } k \text{ sujeito a } \alpha_k > 0 \end{array}}$$

$$\boxed{f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x}_i^{\mathrm{T}} \mathbf{x} + b}$$

SVM não-linear com variável de folga

Resolvendo o problema dual teremos:

$\mathbf{w} = \Sigma \alpha_i y_i \mathbf{x}_i$

$b = y_k(1 - \xi_k) - \Sigma \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ para qualquer k sujeito a $\alpha_k > 0$

$f(\mathbf{x}) = \Sigma \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b$