

# Reconhecimento de padrões – aula 9

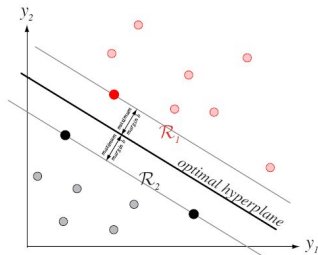
## 11 de Novembro de 2010

Manuel Campagnolo

actualizado em 18 de Novembro de 2010

# Funções discriminantes lineares: Introdução a máquinas de vetores de suporte (SVM)

Ideia: dados dois conjuntos de observações  $S_1$  e  $S_2$ , maximizar a *margem* entre as regiões de decisão.



**FIGURE 5.19.** Training a support vector machine consists of finding the optimal hyperplane, that is, the one with the maximum distance from the nearest training patterns. The support vectors are those (nearest) patterns, a distance  $b$  from the hyperplane. The three support vectors are shown as solid dots. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

As observações (ampliadas)  $\mathbf{y} = (1, x_1, \dots, x_d)$  que verificam a igualdade  $\mathbf{y}^t \mathbf{w} = 0$  situam-se num hiperplano no espaço das observações.

Como na aula anterior (conjuntos linearmente separáveis) procura-se saber se existe um hiperplano que separa as observações de cada grupo.

Mas, para além disso, procura-se o hiperplano que fica o mais afastado possível das observações de cada grupo.

# Observações normalizadas

Procura-se então o vector  $\mathbf{w} \neq 0$  tal que:

$$\begin{aligned} \mathbf{y}^t \mathbf{w} &\geq 1 && \text{se } \mathbf{y} \in S_1 \\ \mathbf{y}^t \mathbf{w} &\leq -1 && \text{se } \mathbf{y} \in S_2. \end{aligned}$$

Como na aula anterior pode normalizar-se as observações (i.e.  $\mathbf{x}$  e  $\mathbf{y} = (1, \mathbf{x})$  são multiplicados por  $-1$  se estão no grupo  $S_2$ ) por forma a poder escrever-se a desigualdade da mesma forma para todas as observações em  $S_1 \cup S_2$ :

$$\mathbf{y}^t \mathbf{w} \geq 1 \text{ para } \mathbf{y} \in S_1 \cup S_2.$$

No estudo da separabilidade linear, usar-se-ão observações normalizadas (que serão simplesmente representadas por  $\mathbf{x}$  ou pelo correspondente vector ampliado  $\mathbf{y}$ , ).

## Margem e vectores de suporte

As observações mais próximas do hiperplano  $\mathbf{y}^t \mathbf{w} = 0$  (hiperplano separador) são as observações  $\mathbf{y}$  que satisfazem  $\mathbf{y}^t \mathbf{w} = 1$ . A distância entre esses dois hiperplanos designa-se por **margem** e é dada por  $\frac{1}{\|\mathbf{w}\|}$ .

De facto, como  $\mathbf{w}$  é normal ao hiperplano  $\mathbf{y}^t \mathbf{w} = 0$ , a distância entre os dois hiperplanos é dada pela distância entre a origem (que pertence ao hiperplano  $\mathbf{y}^t \mathbf{w} = 0$ ) e o ponto  $\frac{\mathbf{w}}{\|\mathbf{w}\|^2}$  (que pertence ao hiperplano  $\mathbf{y}^t \mathbf{w} = 1$ ). Essa distância é a norma de  $\frac{\mathbf{w}}{\|\mathbf{w}\|^2}$  que é  $\frac{1}{\|\mathbf{w}\|}$ .

As observações que pertencem ao hiperplano de equação  $\mathbf{y}^t \mathbf{w} = 1$  são designadas por **vectores de suporte**. A definição do hiperplano separador vai depender apenas dessas observações.

## Formalização do problema: primeira versão

Maximizar a margem ( $\frac{1}{\|\mathbf{w}\|}$ ) é equivalente a minimizar  $\|\mathbf{w}\|$  ou então,

$$(P1) \min \frac{1}{2} \|\mathbf{w}\|^2 \text{ s.a. } \mathbf{y}^t \mathbf{w} \geq 1.$$

Como  $\|\mathbf{w}\|^2 = \frac{1}{2} \sum_{i=0}^d w_i^2$ , este é um problema de *programação quadrática* (PQ). Um problema PQ é um problema de minimização de uma função quadrática das variáveis sujeita a restrições lineares sobre as mesmas variáveis.

Formulado desta forma, o problema pode ser resolvido usando um algoritmo para PQ.

No entanto, é possível alterar a formulação do problema for forma a poder:

- ▶ tratar o problema se possivelmente algumas observações não satisfazem  $\mathbf{y}^t \mathbf{w} \geq 1$ ;
- ▶ generalizar o problema para o caso não linear.

## Formalização do problema: segunda versão

Prova-se que o problema P1 é equivalente ao novo problema de programação quadrática:

$$(P2) \max L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{k,j=1}^n \alpha_k \alpha_j \mathbf{x}_k^t \mathbf{x}_j$$

$$\text{s.a. } \sum_{i=1}^n z_i \alpha_i = 0 \text{ e } \alpha_i \geq 0, i = 1, \dots, n$$

em que  $z_i = 1$  se  $\mathbf{x}_i \in S_1$  e  $z_i = -1$  se  $\mathbf{x}_i \in S_2$ .

O vector  $\mathbf{w} = (w_0, \underline{\mathbf{w}})$  que define o hiperplano separador satisfaz:

- ▶ Os vectores de suporte são  $S = \{\mathbf{x}_i : \alpha_i > 0\}$ ;
- ▶  $\underline{\mathbf{w}} = \sum_{i=1}^n \alpha_i \mathbf{x}_i = \sum_{\mathbf{x}_i \in S} \alpha_i \mathbf{x}_i$ ;
- ▶ Dado um vector de suporte  $\mathbf{x}_s$ ,  $w_0$  é a solução de  $w_0 + \mathbf{x}_s^t \underline{\mathbf{w}} = 1$ .

# Aplicação do critério

1. (Pré-processamento) Dados  $S_1 = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_1}\}$  e  $S_2 = \{\mathbf{x}_{n_1+1}, \dots, \mathbf{x}_n\}$ , definir  $z_j = 1$  se  $\mathbf{x}_j \in S_1$  e  $z_j = -1$  se  $\mathbf{x}_j \in S_2$ . Normalizar os  $\mathbf{x}_j$ , isto multiplicar cada  $\mathbf{x}_j$  por  $z_j$ .
2. Resolver o problema (P2). O output é o vector dos  $\alpha_j$ ;
3. Se o problema tem solução  $\alpha > 0$ ,  $S_1$  e  $S_2$  são linearmente separáveis;
4. A observação  $\mathbf{x}_j$  é um vector de suporte sse  $\alpha_j > 0$  na solução;
5. Calcular a partir dos  $\alpha_j$  o vector  $\mathbf{w}$ ;
6. Regra de decisão: dada uma nova observação  $\mathbf{y} = (1, \mathbf{x})$ , calcular  $\mathbf{y}^t \mathbf{w}$ . Se  $\mathbf{y}^t \mathbf{w} > 0$  associar  $\mathbf{x}$  à classe de  $S_1$ . Caso contrário, associar à classe de  $S_2$ .

# Máquinas de vetores de suporte: classes não linearmente separáveis

Se não existe  $\mathbf{w}$  tal que todas as observações satisfazem  $\mathbf{y}_i^t \mathbf{w} \geq 1$ , então é necessário relaxar essas condições introduzindo novas variáveis  $\xi_i, i = 1, \dots, n$ .

O novo problema de otimização é

$$(P3) \min \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

$$\text{s.a. } \mathbf{y}_i^t \mathbf{w} \geq 1 - \xi_i \text{ e } \xi_i \geq 0, i = 1, \dots, n$$

em que  $C$  é o parâmetro que define a penalização por violar as condições  $\mathbf{y}_i^t \mathbf{w} \geq 1$ .

# Formulação para classes não linearmente separáveis

O problema P3 é equivalente ao seguinte problema:

$$(P4) \max L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{k,j=1}^n \alpha_k \alpha_j \mathbf{x}_k^t \mathbf{x}_j$$

$$\text{s.a. } \sum_{i=1}^n z_i \alpha_i = 0 \text{ e } 0 \leq \alpha_i \leq \frac{C}{n}, i = 1, \dots, n$$

em que  $z_i = 1$  se  $\mathbf{x}_i \in S_1$  e  $z_i = -1$  se  $\mathbf{x}_i \in S_2$  e  $C > 0$  é um parâmetro que traduz o custo de violar as condições de separabilidade linear.

# Aplicação do critério

1. (Pré-processamento) Dados  $S_1 = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_1}\}$  e  $S_2 = \{\mathbf{x}_{n_1+1}, \dots, \mathbf{x}_n\}$ , definir  $z_i = 1$  se  $\mathbf{x}_i \in S_1$  e  $z_i = -1$  se  $\mathbf{x}_i \in S_2$ . Normalizar os  $\mathbf{x}_i$ , isto multiplicar cada  $\mathbf{x}_i$  por  $z_i$ ;
2. Escolher um valor para o parâmetro  $C$ ;
3. Resolver o problema (P4). O output é o vector dos  $\alpha_i$ ;
4. A observação  $\mathbf{x}_i$  é um vector de suporte sse  $0 < \alpha_i \leq \frac{C}{n}$  na solução;
5. Calcular a partir dos  $\alpha_i$  o vector  $\mathbf{w}$ ;
6. Regra de decisão: dada uma nova observação  $\mathbf{y} = (1, \mathbf{x})$ , calcular  $\mathbf{y}^t \mathbf{w}$ . Se  $\mathbf{y}^t \mathbf{w} > 0$  associar  $\mathbf{x}$  à classe de  $S_1$ . Caso contrário, associar à classe de  $S_2$ .

# Máquinas de vetores de suporte não lineares

Ideia:  $S_1$  e  $S_2$  podem não ser linearmente separáveis no espaço de representação das observações mas serem linearmente separáveis num espaço das variáveis transformadas.

Exemplo. Considere  $d = 1$  e as observações  $S_1 = \{-4, -3, -2, 5, 8\}$ ,  $S_2 = \{-1, .5, 1.5, 3\}$ .  $S_1$  e  $S_2$  não são linearmente separáveis no espaço de representação original (variável  $x$ ) mas são linearmente separáveis no espaço das variáveis  $x$  e  $x^2$ .

Em geral, usa-se uma transformação

$$\mathbf{x} = (x_1, \dots, x_d) \rightarrow \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_D(\mathbf{x})).$$

Nota: para este tipo de máquinas, considerar-se-á que as observações  $\mathbf{x}$  *não* estão normalizadas.

## Kernel da transformação

*Kernel trick.* As máquinas de vectores de suporte tem a característica de otimizarem uma função que apenas depende dos termos  $\mathbf{x}_k^t \mathbf{x}_j$  (ver formulações de P2 e de P4). Mostra-se que se pode substituir nas formulações anteriores e nas soluções respectivas os termos  $\mathbf{x}_k^t \mathbf{x}_j$  por  $K(\mathbf{x}_k, \mathbf{x}_j)$ , em que  $K$  verifica  $K(\mathbf{x}_k, \mathbf{x}_j) = \phi(\mathbf{x}_k)^t \phi(\mathbf{x}_j)$  para alguma transformação  $\phi$ . Assim, pode considerar-se o espaço dado pela transformação subjacente  $\phi$  sem ter que considerar (ou computar)  $\phi$  explicitamente.

Alguns exemplos de *kernels*:

1. linear:  $K(\mathbf{x}_k, \mathbf{x}_j) = \mathbf{x}_k^t \mathbf{x}_j$ ;
2. polinomial:  $K(\mathbf{x}_k, \mathbf{x}_j) = (c \mathbf{x}_k^t \mathbf{x}_j + b)^p$ ;
3. gaussiano:  $K(\mathbf{x}_k, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_k - \mathbf{x}_j\|^2)$ ;
4. radial:  $K(\mathbf{x}_k, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_k - \mathbf{x}_j\|)$ ;
5. sigmoide:  $K(\mathbf{x}_k, \mathbf{x}_j) = \tanh(c \mathbf{x}_k^t \mathbf{x}_j + b)$ , com  $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ .

## Formulação do problema no caso geral: possivelmente não linear e não separável

A máquina de vectores de suporte que devolve a solução óptima do seguinte problema designa-se por **C-SVM**.

$$(P5) \quad \max L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{k,j=1}^n \alpha_k \alpha_j z_k z_j K(\mathbf{x}_k, \mathbf{x}_j)$$

$$\text{s.a.} \quad \sum_{i=1}^n z_i \alpha_i = 0 \quad \text{e} \quad 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \dots, n$$

em que  $z_i = 1$  se  $\mathbf{x}_i \in S_1$  e  $z_i = -1$  se  $\mathbf{x}_i \in S_2$  e  $C > 0$  é um parâmetro que traduz o custo de violar as condições de separabilidade no espaço das variáveis transformadas  $\phi(\mathbf{x})$  subjacente ao *kernel*  $K$ .

## Formulação do problema no caso geral: utilização do parâmetro $\nu$ em alternativa a $C$

Um alternativa à formulação anterior é usar um parâmetro  $\nu$  que toma valores entre 0 e 1 e que é um limite superior sobre proporção das observações que violam a margem. A máquina de vectores de suporte designa-se então por  $\nu$ -SVM.

$$\begin{aligned} (\text{P6}) \quad \max L(\alpha) &= -\frac{1}{2} \sum_{k,j=1}^n \alpha_k \alpha_j z_k z_j K(\mathbf{x}_k, \mathbf{x}_j) \\ \text{s.a.} \quad &\sum_{i=1}^n z_i \alpha_i = 0 \\ &\sum_{i=1}^n \alpha_i \geq \nu \\ &0 \leq \alpha_i \leq \frac{1}{n}, \quad i = 1, \dots, n \end{aligned}$$

em que  $z_i = 1$  se  $\mathbf{x}_i \in S_1$  e  $z_i = -1$  se  $\mathbf{x}_i \in S_2$ .

# Aplicação do critério

1. Escolher um *kernel*  $K$  e os valores dos seus parâmetros;
2. Escolher um valor para o parâmetro  $C$ ;
3. Resolver o problema (P5); O output é o vector dos  $\alpha_j$ ;
4. A observação  $\mathbf{x}_s$  é um vector de suporte sse  $0 < \alpha_s \leq \frac{C}{n}$  na solução;
5. Regra de decisão: dada uma nova observação  $\mathbf{x}$ , calcular

$$f(\mathbf{x}) = \sum_{i=1}^n z_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + w_0,$$

em que  $w_0$  é a solução de  $z_s f(\mathbf{x}_s) = 1$  para algum vector de suporte  $\mathbf{x}_s$ .

Se  $f(\mathbf{x}) > 0$  então  $\mathbf{x}$  é associado à classe de  $S_1$ ; caso contrário é associado à classe de  $S_2$ .

## Aplicação R: função `svm` (e1071)

Esta função do R constitui uma interface com *libsvm* (Chang e Ling 2001), que é essencialmente um algoritmo de PQ adaptado a máquinas de suporte de vectores. Permite usar kernels lineares, polinomiais, radiais e sigmoidais, e permite realizar  $C$ -classificação e  $\nu$ -classificação. Também permite fazer classificação quando há mais de dois grupos usando o critério de maioria em todas as comparações entre duas classes (critério ‘um contra um’).

```
> sep2<-data.frame(x1=rep(1:4,4),
x2=rep(1:4,each=4),
cl=factor(c(0,0,0,1,0,0,1,1,0,1,1,1,1,1,1)))
> model<-svm(cl~ ., data=sep2, method=
"C-classification", kernel="linear", cost=10)
> cbind(model$SV,model$coefs)
> pred<-predict(model,newdata=sep2,
decision.values=T)
> attr(pred,"decision.values")
```

## Aplicação R: função `svm` (continuação)

Determine os vectores de suporte e os seus coeficientes para os dados seguintes. Caso os grupos não sejam linearmente separáveis, use um kernel do tipo “radial”. Para o conjunto `sep2a`, considere um kernel linear e custos de 1 e 10. Para o conjunto `sep2b` considere um kernel linear e um kernel radial, e um custo de 10. Comente.

```
> sep2a<-data.frame(x1=rep(1:4,4),
x2=rep(1:4,each=4),
cl=factor(c(0,0,0,1,0,0,1,1,0,0,1,1,1,1,1,1)))
> sep2b<-data.frame(x1=rep(1:4,4),
x2=rep(1:4,each=4),
cl=factor(c(0,0,0,1,0,0,1,1,0,0,1,1,1,0,1,1)))
```

## Aplicação R: função `svm` (continuação)

Considere os dados `sep2c`. Use um kernel do tipo “radial” e valores dos parâmetros “gamma” entre 0.1 e 1.0, e um custo de 10. Compare com os resultados da aplicação da função `tune.svm`.

```
> sep2c<-data.frame(x1=rep(1:4,4),  
x2=rep(1:4,each=4),  
cl=factor(c(0,0,0,1,0,0,1,1,0,0,1,1,1,0,1,1)))  
> tobj=tune.svm(cl~.,data=sep2c,  
gamma=10^seq(-2,1,.5), cost=10^seq(0,2,.25))  
> summary(tobj)
```

## Aplicação R: função `svm` (continuação)

Considere agora os dados `sep3` (3 classes) e determine a classificação que as funções discriminantes calculadas pelas instruções abaixo induzem sobre os dados.

```
> sep3<-data.frame(x1=rep(1:4,4),  
x2=rep(1:4,each=4),  
cl=factor(c(0,0,0,0,0,0,0,0,1,1,2,2,1,1,2,2)))  
> model<-svm(cl~., data=sep3, method=  
"C-classification", kernel="linear", cost=10)  
> cbind(model$SV,model$coefs)
```