# Examen de los temas 1 a 4 de Aprendizaje Automàtico

ETSINF, Universitat Politècnica de València, 02 de diciembre de 2013

**Apellidos:**   **Nombre:**

## Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas.

1 $\boxed{B}$ Deseamos evaluar un sistema de Aprendizaje Automático utilizando un conjunto de datos de entrenamiento que contiene 1000 muestras y la técnica de *exclusión individual* ("Leaving One Out"), obteniéndose un total de 44 errores. Indicar cuál de las afirmaciones siguientes es correcta:

A) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 1000 muestras.
B) La talla de entrenamiento efectiva es de 999 muestras y el error es del 4.4 %
C) La talla de entrenamiento efectiva es de 900 muestras y el error es del 44 %
D) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 900 muestras.

2 $\boxed{C}$ Al aplicar la técnica de descenso por gradiente a una modificación de la función de Widrow y Hoff

$$q_S(\boldsymbol{\theta}) = \frac{1}{2}\left(\sum_{n=1}^{N}\left(\boldsymbol{\theta}^t\boldsymbol{x}_n - y_n\right)^2 + \boldsymbol{\theta}^t\boldsymbol{\theta}\right),$$

el gradiente $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$ en la iteración $\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) - \rho_k \boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$ es

A) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n$

B) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{\theta}(k)^t\boldsymbol{\theta}(k)$

C) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{\theta}(k)$

D) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{x}_n$

3 $\boxed{C}$ Entre las siguientes propiedades de las funciones discriminantes lineales hay una que es falsa:

A) La función discriminante lineal aplicada en un punto devuelve un valor proporcional a la distancia del punto al correspondiente hiperplano separador.
B) La distancia del origen de coordenadas al hiperplano separador asociado a una función discriminante lineal es $\frac{\theta_0}{\|\boldsymbol{\theta}\|}$
C) Un hiperplano separador tiene asociado una única función discriminante lineal canónica
D) Un hiperplano separador tiene asociado un número infinito de funciones discriminantes lineales

4 $\boxed{B}$ Se quiere aplicar la técnica esperanza-maximización a un problema de estimación de máxima verosimilitud en el que no hay variables latentes o ocultas. En este caso ¿Cuál de las afirmaciones siguientes es correcta?

A) En ese caso no se puede aplicar la técnica esperanza-maximización.
B) En ese caso solo se aplica la etapa de maximización y en una iteración acaba.
C) En ese caso solo se aplica la etapa de maximización y hay que iterar hasta que converja.
D) En ese caso solo se aplica la etapa del cálculo de la esperanza.

# Problema 1 (4 puntos; tiempo estimado: 40 minutos)

En una tarea de clasificación de correos electrónicos como spam o no-spam se dispone de un conjunto $S$ de 500 correos *no-spam* (clase $A$) y 300 *spam* (clase $B$).

a) ¿Cuál sería el logaritmo de la verosimilitud de los mensajes de $S$ si las probabilidades a priori $P(A)$ y $P(B)$ fueran iguales?

b) Las probabilidades a priori puede estimarse por máxima verosimilitud a partir de $S$ como: $P(A) = 5/8,\ P(B) = 3/8$. Derivar estas probabilidades mediante la técnica de optimización de los multimplicadores de Lagrange

c) Calcular el logaritmo de la verosimilitud de $S$ según las probabilidades a prioiri obtenidas en b) y compararla con la obtenida en a)

a)
- Modelo: $\mathbf{\Theta}_0 \equiv (p_A, p_B)^t$: $\quad p_A \equiv P(c = A) = 0.5,\ \ p_B \equiv P(c = B) = 0.5$
- El logaritmo (neperiano) de la verosimilitud es

$$\log P(S \mid \mathbf{\Theta}_0) \ = \ \log(\prod_{i=1}^{500} p_A \prod_{j=1}^{300} p_B) \ = \ 500 \cdot \log 0.5 + 300 \cdot \log 0.5 \ = -554.52 \quad \Rightarrow \quad P(S \mid \mathbf{\Theta}_0) \ = \ 1.496 \cdot 10^{-241}$$

b)
- Modelo: $\mathbf{\Theta} \equiv (p_A, p_B)^t$, con $p_A + p_B = 1$
- Verosimilitud y logaritmo de la verosimilitud:

$$\begin{aligned} P(S \mid \mathbf{\Theta}) &= \prod_{i=1}^{500} p_A \prod_{j=1}^{300} p_B = p_A^{500} p_B^{300} \\ L_S(\mathbf{\Theta}) &= \log P(S \mid \mathbf{\Theta}) = 500 \log p_A + 300 \log p_B \end{aligned}$$

- Estimación de máxima verosimilitud:

$$\mathbf{\Theta}^\star \ = \ \arg\max_{\mathbf{\Theta}} L_S(\mathbf{\Theta}) \ = \ \arg\max_{\substack{p_A, p_B \\ p_A + p_B = 1}} \ (500 \log p_A + 300 \log p_B)$$

- Lagrangiana: $\Lambda(p_A, p_B, \beta) = 500 \log p_A + 300 \log p_B + \beta\,(1 - p_A - p_B)$
- Soluciones óptimas en función del multiplicador de Lagrange:

$$\left.\begin{aligned} \frac{\partial \Lambda}{\partial p_A} &= \frac{500}{p_A} - \beta = 0 \\ \frac{\partial \Lambda}{\partial p_B} &= \frac{300}{p_B} - \beta = 0 \end{aligned}\right\} \qquad \begin{aligned} p_A^\star(\beta) &= \frac{500}{\beta} \\ p_B^\star(\beta) &= \frac{300}{\beta} \end{aligned}$$

- Función dual de Lagrange:

$$\Lambda_D(\beta) \ = \ 500 \log \frac{500}{\beta} + 300 \log \frac{300}{\beta} + \beta(1 - \frac{500}{\beta} - \frac{300}{\beta}) \ = \ \beta - 800 \log \beta - 800 + 500 \log 500 + 300 \log 300$$

- Valor óptimo del multiplicador de Lagrange: $\dfrac{d\Lambda_D}{d\beta} \ = \ 1 - \dfrac{800}{\beta} \ = \ 0 \ \Rightarrow \ \beta^\star \ = \ 800$
- Solución final: $\boldsymbol{\theta}^\star \ = \ (p_A^\star, p_B^\star)^t :\qquad p_A^\star = p_A^\star(\beta^\star) = \dfrac{5}{8} \qquad p_B^\star = p_B^\star(\beta^\star) = \dfrac{3}{8}$

c) Como en a):

$$L_S(\mathbf{\Theta}^\star) \ = \ 500 \cdot \log(5/8) + 300 \cdot \log(3/8) \ = -529.25 \quad \Rightarrow \quad P(S \mid \mathbf{\Theta}^\star) = 1.411 \cdot 10^{-230} \ \gg \ 1.496 \cdot 10^{-241} = P(S \mid \mathbf{\Theta}_0)$$

La verosmilitud es mayor que en a) debido a que se ha maximizado la verosimilitud con respecto a $\mathbf{\Theta}$.

## Problema 2 (4 puntos; tiempo estimado: 20 minutos)

Para el aprendizaje de una máquina de vectores soporte se dispone de una muestra de entrenamiento linealmente separable

$$S = \{((1,4),+1),((1,6),+1),((2,2),+1),((2,3),+1),((4,2),-1),((3,4),-1),((3,5),-1),((5,4),-1),((5,6),-1)\}$$

Los multiplicadores de Lagrange óptimos son: $\boldsymbol{\alpha}^* \ = \ (0,\ 0.25,\ 0,\ 1.0,\ 0,\ 1.25,\ 0,\ 0,\ 0)^t$.

- Obtener la función discriminante lineal correspondiente
- Calcular el margen óptimo
- Clasificar la muestra $(4,5)$.

- La función discriminante lineal
  - El vector de pesos:
  $$\begin{aligned} \theta_1^\star &= \ +1\cdot 0.25\cdot 1 \ +1\cdot 1.0\cdot 2 \ -1\cdot 1.25\cdot 3 = -1.5 \\ \theta_2^\star &= \ +1\cdot 0.25\cdot 6 \ +1\cdot 1.0\cdot 3 \ -1\cdot 1.25\cdot 4 = -0.5 \end{aligned}$$
  - El peso umbral (con la muestra 2) $\theta_0^* = (+1) - (-1.5\ 1 - 0.5\ 6) = 5.5$
  - La FDL: $\phi(\boldsymbol{x}) = -1.5\ x_1 - 0.5\ x_2 + 5.5$
- Margen óptimo:
$$\frac{2}{\|\boldsymbol{\theta}^\star\|} \ = \ \frac{2}{\sqrt{0.25 + 1.0 + 0,1.25}} = 1.26$$
- Clasificación de la muestra $(4,5)$: $\phi(4,5)\ = -1.5\cdot 4 - 0.5\cdot 5 + 5.5\ = -3\ <\ 0 \quad \Rightarrow \quad \text{clase} = -1$