Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 7 de enero de 2020

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

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas. Cada acierto suma 1/2 puntos y cada fallo resta 1/6 puntos.

1 $\boxed{\text{A}}$ Se ha evaluado un sistema de Aprendizaje Automático mediante la técnica de *validación cruzada en B bloques* ("B-fold Cross Validation") con $B = 100$ y utilizando un conjunto de datos etiquetados que contiene 500 muestras. Se han obtenido un total de 55 errores. Indicar cuál de las afirmaciones siguientes es razonable:

A) Las tallas de entrenamiento y test efectivas son 495 y 500 muestras, respectivamente, y el error estimado es $11.0 \pm 2.7\,\%$
B) Las tallas de entrenamiento y test efectivas son 100 y 400 muestras, respectivamente, y el error estimado es $13.8 \pm 3.0\,\%$
C) Las tallas de entrenamiento y test efectivas son 5 y 495 muestras, respectivamente y el error estimado es $11.1 \pm 2.8\,\%$
D) Ninguna de las anteriores afirmaciones es razonable

2 $\boxed{\text{A}}$ Considerar la siguiente modificación de la función de Widrow y Hoff

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

Al aplicar la técnica de descenso por gradiente, en la iteración k el vector de pesos, $\boldsymbol{\theta}$, se modifica como: $\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) - \rho_k \boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$. En este caso, ¿cuál de los siguintes gradientes, $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$, es correcto?:

A) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \lambda \frac{\boldsymbol{\theta}(k)}{\boldsymbol{\theta}(k)^t \boldsymbol{\theta}(k)}$
B) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \lambda\ \boldsymbol{\theta}(k)$
C) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \frac{\lambda}{2}$
D) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \lambda \log \boldsymbol{\theta}(k)$

3 $\boxed{\text{A}}$ En el problema de optimización con restricciones

$$\begin{array}{lll} \text{minimizar} & q(\boldsymbol{\Theta}), & \boldsymbol{\Theta} \in \mathbb{R}^D \\ \text{sujecto a} & v_i(\boldsymbol{\Theta}) \leq 0, & 1 \leq i \leq k \end{array}$$

se cumplen las condiciones complementarias de Karush-Kuhn-Tucker $\alpha_i^\star v_i(\boldsymbol{\Theta}^\star) = 0$ para $1 \leq i \leq k$. Indicar cuál de las siguientes afirmaciones se deduce de ellas:

A) Para todo i tal que $\alpha_i^\star > 0$, entonces $v_i(\boldsymbol{\Theta}^\star) \ = \ 0$
B) Para todo i tal que $\alpha_i^\star < 0$, entonces $v_i(\boldsymbol{\Theta}^\star) \ = \ 0$
C) Si para un i, $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) \ = \ 0$
D) Para todo i, si $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$,

4 $\boxed{\text{A}}$ Las siguientes afirmaciones se refieren a la estimación por máxima verosimilitud de los parámetros de una mezcla de K gaussianas (vector-media y peso de cada gaussiana) mediante un conjunto de vectores de entrenamiento cualquiera de dimensión D. Identifica cuál es *falsa*.

A) El algoritmo *esperanza-maximización* es una alternativa a la técnica de los Multiplicadores de Lagrange en el caso de la estimación de los parámetros de una mezcla de K gaussianas.
B) La verosimilitud del conjunto de entrenamiento, calculada con los parámetros estimados, aumenta en cada iteración del *esperanza-maximización*.
C) En cada iteración, el algoritmo *esperanza-maximización* realiza una estimación de los valores de los pesos de las gaussianas.
D) Los parámetros de la mezcla se estiman adecuadamente mediante un algoritmo de *esperanza-maximización*

Problema 1 (3 puntos; tiempo estimado: 30 minutos)

En la siguiente tabla se presenta una muestra de entrenamiento no linealmente separable en $\mathbb{R}^2$ y los correspondientes multiplicadores de Lagrange óptimos obtenidos al entrenar una máquina de vectores soporte con esta muestra (y C=10):

i	1	2	3	4	5	6	7	8
x_{i1}	4	2	4	2	1	2	3	4
x_{i2}	3	5	2	2	4	3	4	4
Clase	–1	–1	+1	+1	+1	–1	–1	–1
$\alpha_i^\star$	6.22	0	9.11	0	7.11	10	0	0

a) Obtener la función discriminante lineal correspondiente

b) Representar gráficamente la frontera lineal de separación entre clases y las muestras de entrenamiento, indicando cuáles son vectores soporte.

c) Clasificar la muestra $(1,1)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_5\alpha_5^\star\ \mathbf{x_5} + c_6\alpha_6^\star\ \mathbf{x_6}$

$$\theta_1^* = (-1)\ (6.22)\ (4) + (+1)\ (9.11)\ (4) + (+1)\ (7.11)\ (1) + (-1)\ (10.0)\ (2) = -1.33$$
$$\theta_2^* = (-1)\ (6.22)\ (3) + (+1)\ (9.11)\ (2) + (+1)\ (7.11)\ (4) + (-1)\ (10.0)\ (3) = -2.00$$

Usando el vector soporte $\mathbf{x_1}$ (que verifica la condición : $0 < \alpha_1^* < C$)

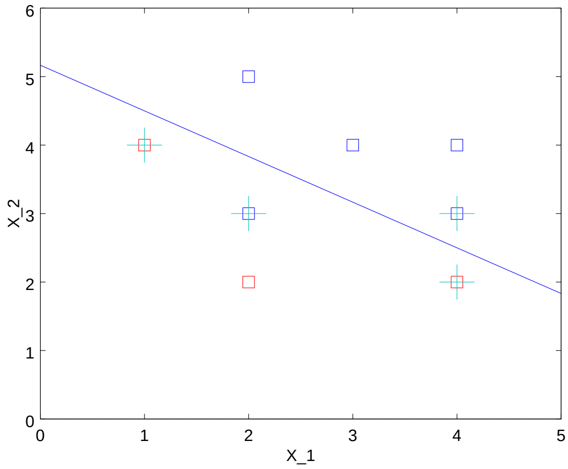
$\theta_0^\star = c_1 - \boldsymbol{\theta}^{\star t}\mathbf{x_1} = -1 - ((-1.33)\ (4) - (2.00)\ (3)) = 10.33$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $10.33 - 1.33\ x_1 - 2.00\ x_2 = 0 \ \rightarrow\ x_2 = -0.665\ x_1 + 5.165\,.$

Los vectores de entrenamiento son todos los de la tabla. De ellos, los vectores soporte son: $(4,3)^t, (4,2)^t, (1,4)^t, (2,3)^t$

Representación gráfica:

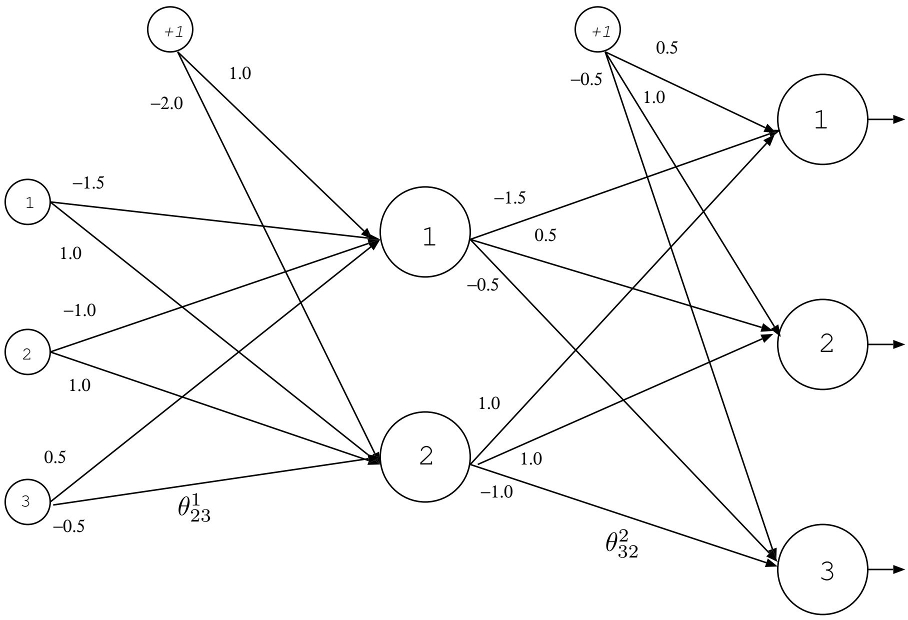


c) Clasificación de la muestra $(1,1)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 5 + \theta_2^*\ 5\ = +7.0 > 0\ \Rightarrow$ clase $+1$.

Problema 2 (3 puntos; tiempo estimado: 30 minutos)

El perceptrón multicapa de la figura se utiliza para resolver un problema de regresión.



Se asume que la función de activación de los nodos de la capa de salida y de la capa oculta es de tipo sigmoid. Sean:

Un vector de entrada : $x_1 = 2.0, \quad x_2 = 1.0, \quad x_3 = 2.0$

Los valores deseados de la capa de salida : $t_1 = 2.0, \quad t_2 = 1.0, \quad t_3 = 2.0$

Calcular:

a) Los valores que se obtienen en la unidades ocultas y de salida.

b) Los correspondientes errores en los tres nodos de la capa de salida y en los dos nodos de la capa oculta.

c) Los nuevos valores de los pesos θ^2_{32} y θ^1_{23} asumiendo que el factor de aprendizaje ρ es 1.0

a) Los valores de la capa de salida son:

$s^1_1 = f_s(\theta^1_{1,0} + \theta^1_{1,1}x_1 + \theta^1_{1,2}x_2 + \theta^1_{1,3}x_3) = 0.1192$; $\quad s^1_2 = f_s(\theta^1_{2,0} + \theta^1_{2,1}x_1 + \theta^1_{2,2}x_2 + \theta^1_{2,3}x_3) = 0.5$

$s^2_1 = f_s(\theta^2_{1,0}+\theta^2_{1,1}s^1_1+\theta^2_{1,2}s^1_2) = 0.6945$; $\quad s^2_2 = f_s(\theta^2_{2,0}+\theta^2_{2,1}s^1_1+\theta^2_{2,2}s^1_2) = 0.8263$; $\quad s^2_3 = f_s(\theta^2_{3,0}+\theta^2_{3,1}s^1_1+\theta^2_{3,2}s^1_2) = 0.2574$

b) Los errores en la capa de salida son:

$\delta^2_1 = (t_1 - s^2_1)\ s^2_1\ (1 - s^2_1) = 0.2770$; $\quad \delta^2_2 = (t_2 - s^2_2)\ s^2_2\ (1 - s^2_2) = 0.0249$; $\quad \delta^2_3 = (t_3 - s^2_3)\ s^2_3\ (1 - s^2_3) = 0.3331$

Los errores en la capa de oculta son:

$\delta^1_1 = (\delta^2_1\ \theta^2_{11} + \delta^2_2\ \theta^2_{21} + \delta^2_3\ \theta^2_{31})\ s^1_1\ (1 - s^1_1) = -0.0598$; $\quad \delta^1_2 = (\delta^2_1\ \theta^2_{12} + \delta^2_2\ \theta^2_{22} + \delta^2_3\ \theta^2_{32})\ s^1_2\ (1 - s^1_2) = -0.0078$

c) El nuevo peso θ^2_{32} es: $\quad \theta^2_{32} = \theta^2_{32} + \rho\ \delta^2_3\ s^1_2 = (-1.0) + (1)\ (0.3331)\ (0.5) = -0.8335$

El nuevo peso θ^1_{23} es: $\quad \theta^1_{23} = \theta^1_{23} + \rho\ \delta^1_2\ x_3 = (-0.5) + (1)\ (-0.0078)\ (2.0) = -0.5156$

Problema 3 (2 puntos; tiempo estimado: 20 minutos)

Considerar la red bayesiana $\mathcal{R}$ definida como $P(A,B,C,D) = P(A)\ P(B \mid A)\ P(C \mid B)\ P(D \mid B)$, cuyas variables A, B, C, y D toman valores en el conjunto $\{0,1\}$ y sus distribuciones de probabilidad asociadas son:

$$\begin{aligned}
P(A=1) &= 0.3 & P(A=0) &= 0.7 \\
P(B=1 \mid A=1) &= 0.4 & P(B=0 \mid A=1) &= 0.6 \\
P(B=1 \mid A=0) &= 0.6 & P(B=0 \mid A=0) &= 0.4 \\
P(C=1 \mid B=1) &= 0.2 & P(C=0 \mid B=1) &= 0.8 \\
P(C=1 \mid B=0) &= 0.7 & P(C=0 \mid B=0) &= 0.3 \\
P(D=1 \mid B=1) &= 0.1 & P(D=0 \mid B=1) &= 0.9 \\
P(D=1 \mid B=0) &= 0.5 & P(D=0 \mid B=0) &= 0.5
\end{aligned}$$

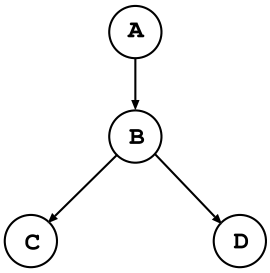
a) Representar gráficamente la red

b) Obtener una expresión simplificada de $P(B,C,D \mid A)$ y calcular su valor para $B=1, C=1$ y $D=1$ cuando $A=0$.

c) Obtener una expresión simplificada de $P(B \mid A,C,D)$ en función de las distribuciones definidas en los nodos de $\mathcal{R}$ y calcular su valor para $B=0$ cuando $A=1, C=1$ y $D=1$.

d) Dados $A=1, C=1$ y $D=1$, ¿Cuál es la mejor predicción para el valor de B?

a) Representación gráfica de la red:



b) Obtener una expresión simplificada de $P(B,C,D \mid A)$ y calcular su valor para $B=1, C=1$ y $D=1$ cuando $A=0$.

$$P(B,C,D \mid A) = \frac{P(A,B,C,D)}{P(A)} = P(B \mid A)\ P(C \mid B)\ P(D \mid B)$$

$$P(B=1, C=1, D=1 \mid A=0) = 0.6 \cdot 0.2 \cdot 0.1 = 0.012$$

c) Obtener una expresión simplificada de $P(B \mid A,C,D)$ en función de las distribuciones definidas en los nodos de $\mathcal{R}$ y calcular su valor para $B=0$ cuando $A=1, C=1$ y $D=1$.

$$\begin{aligned}
P(B \mid A,C,D) &= \frac{P(A,B,C,D)}{P(A,C,D)} \\
&= \frac{P(A)\ P(B \mid A)\ P(C \mid B)\ P(D \mid B)}{\sum_b P(A)\ P(B=b \mid A) P(C \mid B=b) P(D \mid B=b)} \\
&= \frac{P(B \mid A)\ P(C \mid B)\ P(D \mid B)}{\sum_b P(B=b \mid A) P(C \mid B=b) P(D \mid B=b))}
\end{aligned}$$

$$P(B=0 \mid A=1, C=1, D=1) = \frac{0.6 \cdot 0.7 \cdot 0.5}{0.6 \cdot 0.7 \cdot 0.5 + 0.4 \cdot 0.2 \cdot 0.1} = 0.9633$$

d) Dados $A=1, C=1$ y $D=1$, ¿Cuál es la mejor predicción para el valor de B?

$b^\star = \arg\max_{b \in \{0,1\}} P(B=b \mid A=1, C=1, D=1)$

$P(B=1 \mid A=1, C=1, D=1) = 1 - 0.9633 = 0.0367$, por tanto la mejor predicción es $B=0$