Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 16 de enero de 2023

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 $\boxed{\text{C}}$ 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 de que contiene 1000 muestras. Se han obtenido un total de 22 errores. Indicar cuál de las afirmaciones siguientes es razonable:

A) La talla de entrenamiento efectiva es 990 muestras y el error estimado es $2.2\,\% \pm 0.1\,\%$
B) La talla de entrenamiento efectiva es de 900 muestras y el error estimado es $2.2\,\%$
C) La talla de test efectiva es de 1000 muestras y el error estimado es $2.2\,\% \pm 0.9\,\%$
D) El error estimado es $22\,\% \pm 9\,\%$.

2 $\boxed{\text{A}}$ Considerar el aprendizaje mediante máquinas de vectores soportes y márgenes blandos con una muestra de aprendizaje $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ no separable linealmente. Si un multiplicador de Lagrange óptimo α_j^*, asociado a la restricción $c_j\ (\boldsymbol{\theta}^t \boldsymbol{x}_j + \theta_0)\ \geq\ 1 - \zeta_j$, $1 \leq j \leq N$, es cero, indicar la respuesta correcta:

A) La muestra $\boldsymbol{x}_j$ está clasificada correctamente.
B) La muestra $\boldsymbol{x}_j$ está mal clasificada.
C) La muestra $\boldsymbol{x}_j$ está clasificada correctamente pero $\boldsymbol{\theta}$ y θ_0 no es canónico con respecto a la muestra.
D) La muestra $\boldsymbol{x}_j$ es un vector soporte.

3 $\boxed{\text{C}}$ Considerar la siguiente modificación de la función de Widrow y Hoff

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

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 esta expresión, el gradiente, $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$, es:

A) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + 1$
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) $\displaystyle\sum_{n=1}^{N} \boldsymbol{\theta}(k)^t \boldsymbol{x}_n + 1$

4 $\boxed{\text{A}}$ Sea $\mathcal{C}$ un conjunto de variables aleatorias. Un concepto importante en el que se basan las técnicas de redes bayesianas es:

A) El grafo que relaciona a las variables entre si define una distribución de probabilidad conjunta en las variables $\mathcal{C}$ y permite calcular cualquier probabilidad condicional en la que intervengan variables de $\mathcal{C}$.
B) Los nodos del grafo representan las dependencias entre las variables en $\mathcal{C}$.
C) El grafo que relaciona a las variables entre si define una distribución de probabilidad condicional entre dos subconjuntos de variables en $\mathcal{C}$.
D) Las probabilidades condicionales se calculan a partir de los cliques (subgrafos completos) que contiene el grafo.

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

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 $(5, 5)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_4\alpha_4^\star\ \mathbf{x_4} + c_8\alpha_8^\star\ \mathbf{x_8}$

$$\theta_1^* = (+1)\ (1)\ (7.11) + (-1)\ (2)\ (10) + (+1)\ (4)\ (9.11) + (-1)\ (4)\ (6.22) = -1.33$$
$$\theta_2^* = (+1)\ (4)\ (7.11) + (-1)\ (3)\ (10) + (+1)\ (2)\ (9.11) + (-1)\ (3)\ (6.22) = -2.00$$

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

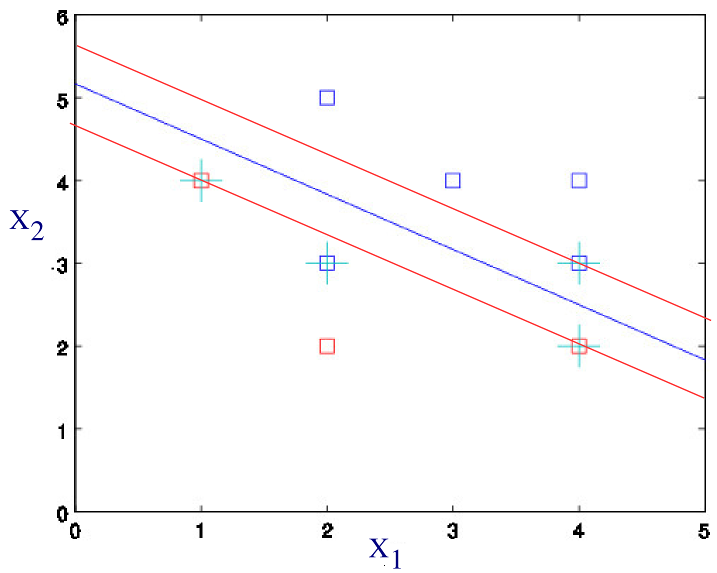
$\theta_0^\star = c_1 - \boldsymbol{\theta}^{\star t}\mathbf{x_1} = 1 - ((-1.33)\ (1) - (2.00)\ (4)) = 10.33$

b) Frontera de separación y representación gráfica:

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

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

Representación gráfica:

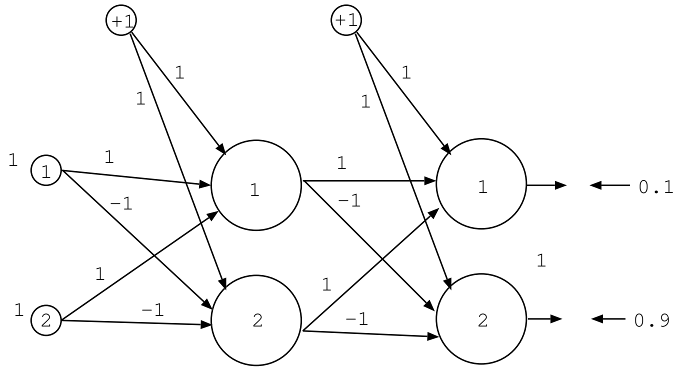


c) Clasificación de la muestra $(5, 5)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 5 + \theta_2^*\ 5 = -6.32 < 0 \Rightarrow$ clase -1.

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

En la red de la figura, para un resolver un problemas de regresión, se utilizan funciones de activación de tipo *sigmoid* en los nodos de la capa de salida y de la capa oculta y como factor de aprendizaje se ha escogido $\rho = 1.0$.



Dados los pesos iniciales indicados en la figura, un vector de entrada $\boldsymbol{x}^t = (1, 1)$ y su valor deseado de salida $t = (0.1, 0.9)$, Calcular:

a) las salidas de todas las unidades

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

c) Los nuevos valores de los pesos de las conexiones al nodo 2 de la capa oculta.

a) Las salidas de todas las unidades

Capa oculta

$$\phi_1^1 = \theta_{10}^1 + \theta_{11}^1\ x_1 + \theta_{12}^1\ x_2 = 3 \qquad s_1^1 = \frac{1}{1 + \exp\left(-\phi_1^1\right)} = 0.953$$

$$\phi_2^1 = \theta_{20}^1 + \theta_{21}^1\ x_1 + \theta_{22}^1\ x_2 = -1 \qquad s_2^1 = \frac{1}{1 + \exp\left(-\phi_2^1\right)} = 0.269$$

Capa de salida

$$\phi_1^2 = \theta_{10}^2 + \theta_{11}^2\ s_1^1 + \theta_{12}^2\ s_2^1 = 2.221 \qquad s_1^2 = \frac{1}{1 + \exp\left(-\phi_1^2\right)} = 0.902$$

$$\phi_2^2 = \theta_{20}^2 + \theta_{21}^2\ s_1^1 + \theta_{22}^2\ s_2^1 = -0.222 \qquad s_2^2 = \frac{1}{1 + \exp\left(-\phi_2^2\right)} = 0.445$$

b) Los errores en la capa de salida son:

$\delta_1^2 = (t_1 - s_1^2)\ s_1^2\ (1 - s_1^2) = -0.0708 \qquad \delta_2^2 = (t_2 - s_2^2)\ s_2^2\ (1 - s_2^2) = +0.1124$

Los errores en la capa de oculta son:

$\delta_1^1 = (\delta_1^2\ \theta_{11}^2 + \delta_2^2\ \theta_{21}^2)\ s_1^1\ (1 - s_1^1) = -0.0082 \qquad \delta_2^1 = (\delta_1^2\ \theta_{12}^2 + \delta_2^2\ \theta_{22}^2)\ s_2^1\ (1 - s_2^1) = -0.0360$

c) Los nuevos pesos del nodo 2 son:

$\theta_{20}^1 = \theta_{20}^1 + \rho\ \delta_2^1\ (+1) = 0.964$

$\theta_{21}^1 = \theta_{21}^1 + \rho\ \delta_2^1\ x_1 = -1.036$

$\theta_{22}^1 = \theta_{22}^1 + \rho\ \delta_2^1\ x_2 = -1.036$

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

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

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

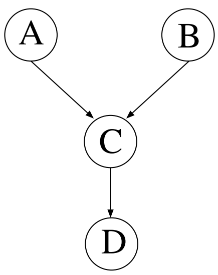
a) Representar gráficamente la red

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

c) Dados $B=1, C=1$ y $D=1$, ¿Cuál es el valor óptimo de A?

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

a) Representación gráfica de la red:



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

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

$$P(A=0 \mid B=1,C=1,D=1) = \frac{0.7 \cdot 0.2}{0.7 \cdot 0.2\ +\ 0.3 \cdot 0.4} = 0.5385$$

c) Dados $B=1, C=1$ y $D=1$, ¿Cuál es el valor óptimo de A?

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

$P(A=1 \mid B=1,C=1,D=1) = 1 - 0.5385 = 0.4615$, por tanto el valor óptimo es $A=0$

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

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

$$P(B=1,C=1,D=1 \mid A=0) = 0.4 \cdot 0.2 \cdot 0.7 = 0.056$$