Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 7 de enero de 2019

Apellidos: _____ **Nombre:** _____

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas. Cada acierto suma 0.4 puntos y cada fallo resta 1/6 puntos.

1 [D] Al aplicar el método de partición de datos denominado validación cruzada en 5 bloques a un conjunto de 1000 muestras, el clasificador obtiene, por bloque, 2, 5, 7, 8, 5 errores. Indicar la opción *correcta*:

A) El error es de $0.5\,\% \pm 0.5\,\%$.
B) La talla de entrenamiento efectiva es de 900 muestras.
C) La talla de entrenamiento efectiva es de 1000 muestras.
D) El error es de $2.7\,\% \pm 1\,\%$.

2 [C] En un clasificador en 3 clases resulta que la probabilidad a-posteriori de cada clase, y, dada una muestra $\mathbf{x}$ es:

y	$P(Y = y \mid \mathbf{x})$
A	0.1
B	0.6
C	0.3

Indicar cuál es la opción *errónea*:

A) La probabilidad de error si se toma la decisión $Y = B$ es 0.4.
B) La mínima probabilidad de error es 0.4.
C) La probabilidad de error si se toma la decisión $Y = C$ es 0.4.
D) La peor decisión es $Y = A$, cuya probabilidad de error es 0.9.

3 [A] En el problema de optimización con restricciones

$$\begin{array}{ll} \text{minimizar} & q(\boldsymbol{\Theta}), \quad \boldsymbol{\Theta} \in \mathbb{R}^D \\ \text{sujeto a} & v_i(\boldsymbol{\Theta}) \geq 0, \quad 1 \leq i \leq k \\ & u_i(\boldsymbol{\Theta}) = 0, \quad 1 \leq i \leq m \end{array}$$

sea $\boldsymbol{\Theta}^\star$ la solución óptima y sean $\alpha_i^\star$, $1 \leq i \leq k$, y $\beta_i^\star$, $1 \leq i \leq m$, los multiplicadores de Lagrange óptimos para las restricciones de desigualdad e igualdad, respectivamente. Indicar cuál de las siguientes afirmaciones es *falsa*:

A) Si para algún j, $\alpha_j^\star = 0$, entonces $v_j(\boldsymbol{\Theta}^\star) = 0$.
B) Para $1 \leq i \leq m$ $u_i(\boldsymbol{\Theta}^\star) = 0$.
C) Para $1 \leq i \leq k$ $v_i(\boldsymbol{\Theta}^\star) \geq 0$.
D) $v_j(\boldsymbol{\Theta}^\star) = 0$ $\forall j$ si $\alpha_j^\star > 0$, $1 \leq j \leq k$.

4 [D] Las siguientes afirmaciones se refieren a la estimación por máxima verosimilitud de los parámetros de un modelo mediante el algoritmo de *esperanza maximización* (EM). Identificar cuál es *falsa*.

A) En el paso E se estiman los valores de las variables ocultas (o se calculan sus probabilidades).
B) En el paso M se calcula el máximo de una función auxiliar que depende de las estimaciones del paso E.
C) El algoritmo EM se puede aplicar incluso cuando hay restricciones en los valores de los valores de los parámetros o de las variables ocultas, pero para ello hay que aplicar también la técnica de los multiplicadores de Lagrange.
D) Si se usa de algoritmo EM es innecesaria la aplicación de la técnica de los multiplicadores de Lagrange.

5 [C] En una red bayesiana, sea $\mathcal{A}$ un conjunto de variables aleatorias y G el grafo que establece las dependencias entre las variables de $\mathcal{A}$. Identificar cuál de las siguientes afirmaciones es *cierta*.

A) Los arcos de G representan las probabilidades condicionales de las variables de $\mathcal{A}$.
B) G define una distribución de probabilidad condicional entre las variables en $\mathcal{A}$.
C) Cualquier distribución condicional o conjunta en la que participen todas o cualquier subconjunto de las variables de $\mathcal{A}$, se puede deducir a partir de la distribución conjunta definida por G.
D) Si el valor de la variable asociada a un nodo ν de G está dada, entonces todos las variables asociadas a los nodos que están directamente conectados con ν son independientes entre si.

Problema 1 (3 puntos; tiempo estimado: 20 minutos)

Para entrenar un modelo basado en máquinas de vectores soporte, se dispone de un conjunto de entrenamiento en $\mathbb{R}^2$. Estos vectores y los correspondientes multiplicadores de Lagrange óptimos obtenidos con $C = 10$ son:

i	1	2	3	4	5	6	7	8
x_{i1}	1	2	2	2	2	3	4	3
x_{i2}	4	1	2	3	4	2	2	1
Clase	+1	−1	+1	-1	+1	−1	−1	−1
$\alpha_i^\star$	0	3.11	10.0	10.0	3.78	0.67	0	0

a) Obtener la función discriminante lineal correspondiente

b) Obtener la ecuación de la frontera lineal de separación entre clases y representarla gráficamente junto con los vectores de entremamiento, indicando cuáles de ellos son vectores soporte.

c) Obtener la toleranacia óptima de cada muestra de entrenamiento.

d) Clasificar la muestra $(4, 3)^t$.

a) Pesos de la función discriminante:

$$\theta_1^* = +(-1)\ (2)\ (3.11) + (+1)\ (2)\ (10.0) + (-1)\ (2)\ (10.0) + (+1)\ (2)\ (3.79) + (-1)\ (3)\ (0.67) = -0.67$$
$$\theta_2^* = +(-1)\ (1)\ (3.11) + (+1)\ (2)\ (10.0) + (-1)\ (3)\ (10.0) + (+1)\ (4)\ (3.79) + (-1)\ (2)\ (0.67) = 0.67$$

Usando el vector soporte $\mathbf{x_2}$ (que verifica la condición : $0 < \alpha_1^* < C$)

$\theta_0^\star = c_2 - \boldsymbol{\theta}^{\star t}\mathbf{x_2} = 1 - ((-0.667)\ (2) + (0.666)\ (1)) = -0.33$

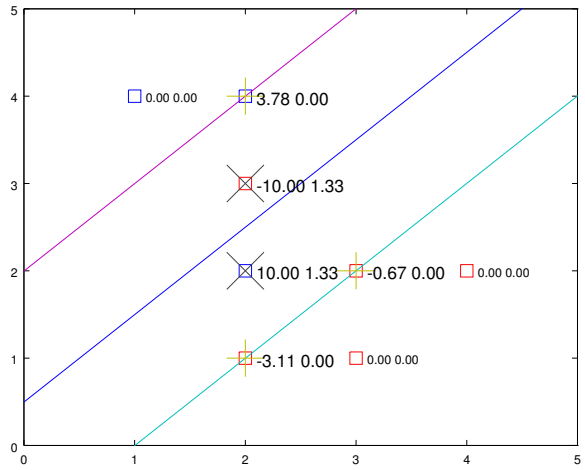
Función discriminante lineal: $\phi(\mathbf{x}) = -0.33 - 0.67\ x_1 + 0.67\ x_2$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $-0.33 - 0.67\ x_1 + 0.67\ x_2 = 0 \rightarrow x_2 = 1.0\ x_1 + 0.49$.

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

Representación gráfica:



Al lado de cada muestra se muestra el valor del multiplicador de lagrange asociado y la tolerancia.

c) Todas las muestras bien clasificadas y fuera del margen ($i \in \{1, 2, 5, 6, 7, 8\}$) tienen una tolerancia $\zeta_i^* = 0$ y el resto

$\zeta_3^* = 1 - c_3\ (\boldsymbol{\theta}^{*t}\ \boldsymbol{x}_3 + \theta_0^*) = 1.33$; $\quad \zeta_4^* = 1 - c_4\ (\boldsymbol{\theta}^{*t}\ \boldsymbol{x}_4 + \theta_0^*) = 1.33$

d) Clasificación de la muestra $(4, 3)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + 4\,\theta_1^* + 3\,\theta_2^* = -1.0 < 0 \Rightarrow$ clase -1.

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

La solución para un determinado problema de regresión viene dado por un perceptrón multicapa, donde la función de activación de todos los nodos de la red son de tipo sigmoid y los pesos en una iteración dada del algoritmo BackProp son:

$$\boldsymbol{\theta}_1^1 = (1.0, -1.0, 0.0, 1.0)^t \quad \boldsymbol{\theta}_2^1 = (-1.0, 0.0, 1.0, -1.5)^t \quad \boldsymbol{\theta}_1^2 = (0.0, -1.0, 0.0)^t \quad \boldsymbol{\theta}_2^2 = (1.0, 1.0, 1.0)^t$$

Supongamos que se dan la circunstancias siguientes:

Un vector de entrada	$x_1 = 0.0$	$x_2 = 1.0$	$x_3 = -1.0$
Las salidas de la capa oculta	$s_1^1 = 0.5$	$s_2^1 = 0.818$	
Las salidas de la capa de salida	$s_1^2 = 0.378$	$s_2^2 = 0.910$	
Los valores deseados de la capa de salida	$t_1 = 0.9$	$t_2 = 0.2$	

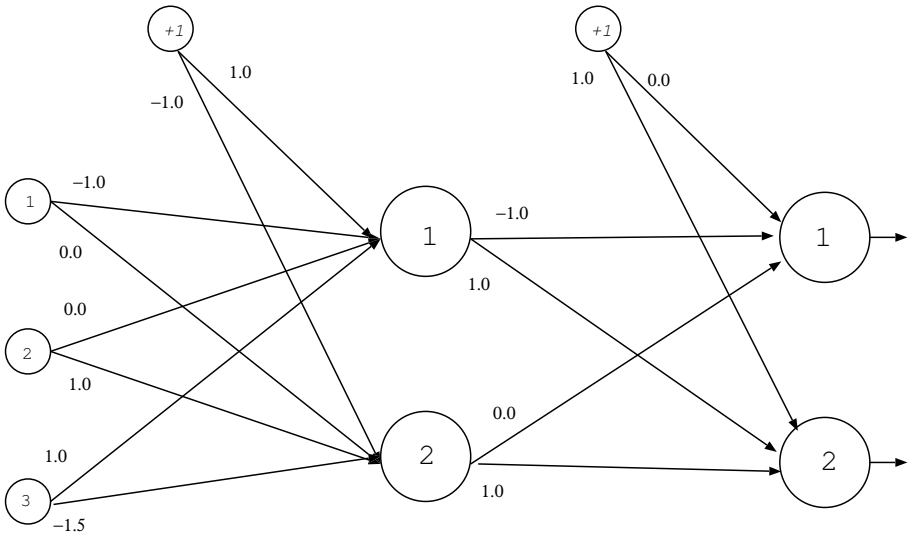
Se pide:

a) Dibujar el perceptrón multicapa descrito al principio del enunciado.

b) Calcular los errores (δ's) en los nodos de la capa de salida y en los nodos de la capa oculta.

c) Calcular los nuevos valores de los pesos $\theta_{2,2}^2$ y $\theta_{2,3}^1$ asumiendo que el factor de aprendizaje ρ es 1.0

a) Dibujo del perceptrón multicapa



b) Errores (δ's) en la capa de salida:

$\delta_1^2 = (t_1 - s_1^2)\ s_1^2\ (1 - s_1^2) = 0.123$
$\delta_2^2 = (t_2 - s_2^2)\ s_2^2\ (1 - s_2^2) = -0.058$

Errores en la capa de oculta:

$\delta_1^1 = (\delta_1^2\ \theta_{11}^2 + \delta_2^2\ \theta_{21}^2) s_1^1\ (1 - s_1^1) = -0.045$
$\delta_2^1 = (\delta_1^2\ \theta_{12}^2 + \delta_2^2\ \theta_{22}^2)\ s_2^1\ (1 - s_2^1) = -0.0086$

c) Nuevo peso $\theta_{2,2}^2 = \theta_{2,2}^2 + \rho\ \delta_2^2\ s_2^1\ =\ 0.953$

Nuevo peso $\theta_{2,3}^1 = \theta_{2,3}^1 + \rho\ \delta_2^1\ x_3\ =\ -1.491$

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

Considerar la red bayesiana $\mathcal{R}$ definida como $P(R,X,Y,Z) = P(X)\ P(R \mid X)\ P(Y \mid R)\ P(Z \mid R)$, cuya variable R toma valores en $\{1,2,3\}$ y las variables X, Y, Z, en el conjunto $\{\texttt{"a"},\texttt{"b"},\texttt{"c"}\}$. Las distribuciones de probabilidad asociadas son como sigue:

- $P(X = \texttt{"a"}) = P(X = \texttt{"b"}) = 1/8$, $P(X = \texttt{"c"}) = 3/4$
- $P(R \mid X)$ es uniforme: $P(R = 1 \mid x) = P(R = 2 \mid x) = P(R = 3 \mid x)$, $\forall x \in \{\texttt{"a"},\texttt{"b"},\texttt{"c"}\}$
- $P(Y \mid R)$ y $P(Z \mid R)$ son idénticas y vienen dadas en la siguiente tabla

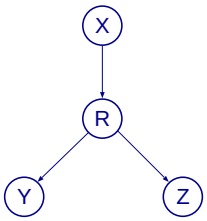
R	"a"	"b"	"c"
1	1/3	0	2/3
2	1/4	1/2	1/4
3	0	3/5	2/5

a) Representar gráficamente la red

b) Obtener una expresión simplificada de $P(X,Y,Z \mid R)$ en función de las distribuciones que definen $\mathcal{R}$

c) calcular $P(X = \texttt{"a"}, Y = \texttt{"a"}, Z = \texttt{"a"} \mid R = 2)$

a) Representación gráfica de la red:



b) Expresión simplificada de $P(X,Y,Z \mid R)$:

$$P(X,Y,Z \mid R) = \frac{P(R,X,Y,Z)}{P(R)} = \frac{P(X)\ P(R \mid X)\ P(Y \mid R)\ P(Z \mid R)}{P(R)}$$

Calculemos el denominador:

$$P(R) = \sum_{xyz} P(R,X,Y,Z) = \sum_x \sum_y \sum_z P(x)\ P(R \mid x)\ P(y \mid R)\ P(z \mid R)$$

$$= \sum_x P(x)P(R \mid x) \sum_y P(y \mid R) \sum_z P(z \mid R) = \Big(\sum_x P(x)P(R \mid x)\Big) \cdot 1 \cdot 1 = \frac{1}{3}\sum_x P(x) = \frac{1}{3}$$

Como $P(R \mid x) = 1/3$ para todo x, resulta $P(X,Y,Z \mid R) = P(X)\ P(Y \mid R)\ P(Z \mid R)$.

c) $P(R \mid X = x) = 1/3\ \forall x \Rightarrow P(X = \texttt{"a"}, Y = \texttt{"a"}, Z = \texttt{"a"} \mid R = 2) = \frac{1}{8} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{128}$