Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 24 de enero de 2018

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 [D] En el problema de optimización con restricciones

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

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) Existe un i tal que $\alpha_i^\star < 0$ y $v_i(\boldsymbol{\Theta}^\star) = 0$
B) Si para algún j, $\alpha_j^\star = 0$, entonces $v_j(\boldsymbol{\Theta}^\star) = 0$
C) Existe un j tal que $v_j(\boldsymbol{\Theta}^\star) > 0$ y $\alpha_j^\star = 0$
D) $v_j(\boldsymbol{\Theta}^\star) = 0$ $\forall j$ tal que $\alpha_j^\star > 0$, $1 \leq j \leq k$

2 [D] 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) Los parámetros de la mezcla se estiman adecuadamente mediante un algoritmo de *esperanza maximización* (EM)
B) En cada iteración, el algoritmo EM estima los valores de las variables ocultas que, en este caso, son los pesos de las gaussianas.
C) La verosimilitud del conjunto de entrenamiento, calculada con los parámetros estimados no disminuye en cada iteración del EM.
D) El algoritmo EM obtiene los valores máximos de los parámetros a estimar.

3 [C] Considerar la siguiente modificación de la función de Widrow y Hoff

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

Cual de las siguientes expresiones del gradiente con respecto a $\boldsymbol{\theta}$ es correcta:

A) $\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) = \sum_{n=1}^{N} \boldsymbol{x}_n$
B) $\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) = \boldsymbol{\theta}^t \sum_{n=1}^{N} \boldsymbol{x}_n$
C) $\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) = \sum_{n=1}^{N} \boldsymbol{x}_n + \lambda\, \boldsymbol{\theta}$
D) $\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) = \sum_{n=1}^{N} \boldsymbol{x}_n + \lambda\, \boldsymbol{\theta}^t \boldsymbol{\theta}$

4 [C] Sea $\mathcal{A}$ un conjunto de variables aleatorias y G el grafo que establece las dependencias entre las variables de $\mathcal{A}$. Un concepto importante en el que se basan las técnicas de redes bayesianas es:

A) Los nodos del G representan las probabilidades incondicionales de las variables de $\mathcal{A}$
B) G define una distribución de probabilidad condicional entre dos subconjuntos de variables en $\mathcal{A}$
C) G define una distribución de probabilidad conjunta de todas las variables de $\mathcal{A}$. A partir de esta distribución, por inferencia probabilística puede calcularse cualquier probabilidad condicional o incondicional en la que intervengan dichas variables
D) G define una distribución de probabilidad conjunta de todas las variables en $\mathcal{A}$. Para calcular las probabilidades de dicha distribución es necesario aplicar reglas de inferencia probabilística tales como la regla de Bayes y la marginalización.

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

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) Clasificar la muestra $(2, 4)^t$.

a) Pesos de la función discriminante:

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

Usando el vector soporte $\mathbf{x_3}$ (que verifica la condición : $0 < \alpha_1^* < C$)

$\theta_0^\star = c_7 - \boldsymbol{\theta}^{\star t}\mathbf{x_3} = 1 - ((-1.33)\ (1) - (2.00)\ (4)) = 10.33$

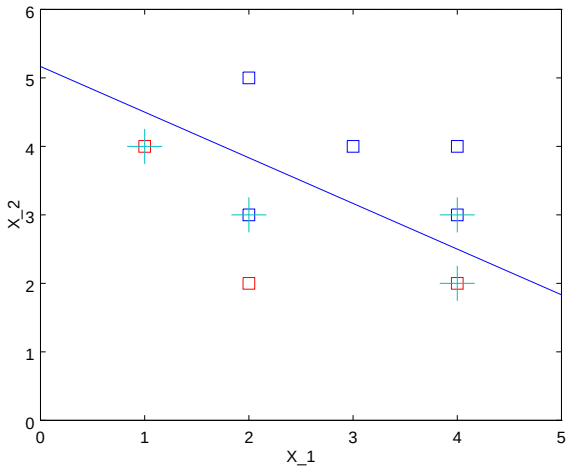
Función discriminante lineal: $\phi(\mathbf{x}) = 10.33 - 1.33\ x_1 - 2.00\ x_2$

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: $(1, 4)^t, (2, 3)^t, (4, 2)^t, (4, 3)^t$.

Representación gráfica:

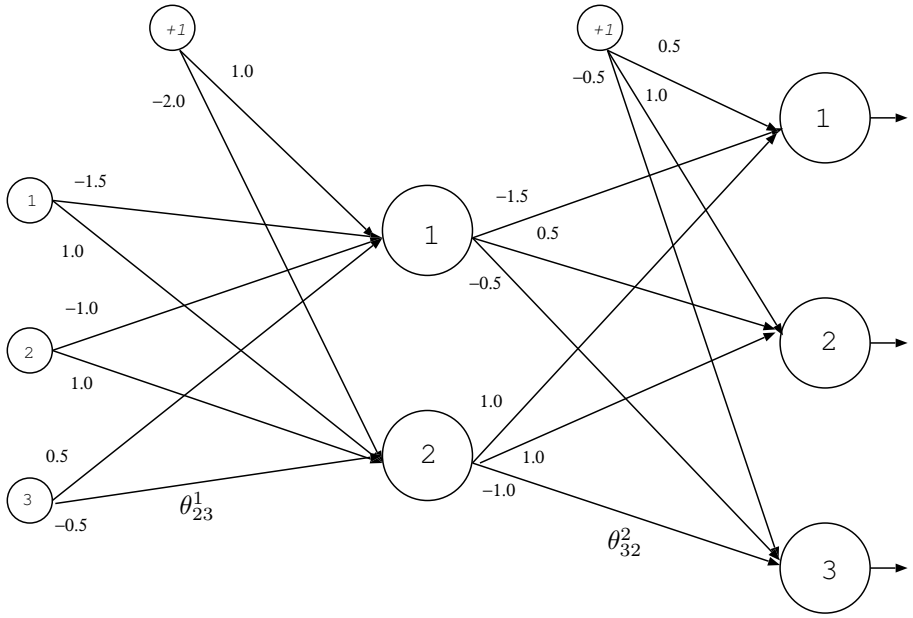


c) Clasificación de la muestra $(2, 4)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + 2\,\theta_1^* + 4\,\theta_2^* = -0.33 < 0 \Rightarrow$ clase -1.

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

La solución para un determinado problema de regresión viene dado por el perceptrón multicapa de la figura, donde las función de activación de todos los nodos de la red son de tipo sigmoid.



Supongamos que se dan la circunstancias siguientes:

Un vector de entrada	$x_1 = 1.0$	$x_2 = 0.0$	$x_3 = 2.0$
Las salidas de la capa oculta	$s^1_1 = 0.622$	$s^1_2 = 0.119$	
Las salidas de la capa de salida	$s^2_1 = 0.442$	$s^2_2 = 0.807$	$s^2_3 = 0.283$
Los valores deseados de la capa de salida	$t_1 = 0.5$	$t_2 = 0.9$	$t_3 = 0.1$

Se pide calcular:

a) Los errores (δ's) en los tres nodos de la capa de salida y en los dos nodos de la capa oculta.

b) Los nuevos valores de los pesos θ^2_{32} y θ^1_{23} asumiendo que el factor de aprendizaje ρ es 2.0

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

$\delta^2_1 = (t_1 - s^2_1)\; s^2_1\; (1 - s^2_1) = (0.5 - 0.442)\, 0.442\, (1 - 0.442) = 0.0143$
$\delta^2_2 = (t_2 - s^2_2)\; s^2_2\; (1 - s^2_2) = (0.9 - 0.807)\, 0.807\, (1 - 0.807) = 0.0145$
$\delta^2_3 = (t_3 - s^2_3)\; s^2_3\; (1 - s^2_3) = (0.1 - 0.283)\, 0.283\, (1 - 0.283) = -0.0371$

Errores en la capa de oculta:

$\delta^1_1 = (\delta^2_1\; \theta^2_{11} + \delta^2_2\; \theta^2_{21} + \delta^2_3\; \theta^2_{31})\; s^1_1\; (1 - s^1_1) = (0.0143(-1.5) + 0.0145 \cdot 0.5 - 0.0371\,(-0.5))\, 0.622\, (1 - 0.622) = 0.0102$
$\delta^1_2 = (\delta^2_1\; \theta^2_{12} + \delta^2_2\; \theta^2_{22} + \delta^2_3\; \theta^2_{32})\; s^1_2\; (1 - s^1_2) = (0.0143 \cdot 1.0 \quad + 0.0145 \cdot 1.0 - 0.0371\,(-1.0))\, 0.119\, (1 - 0.119) = 0.0069$

b) Nuevo peso $\theta^2_{32} = \theta^2_{32} + \rho\; \delta^2_3\; s^1_2 \;=\; (-1.0) + 2\,(-0.0371)\, 0.119 \;= -1.0088$

Nuevo peso $\theta^1_{23} = \theta^1_{23} + \rho\; \delta^1_2\; x_3 \;=\; (-0.5) + 2 \cdot 0.0069 \cdot 2.0 \;= -0.4724$

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

Sean las variables A, B, C, y D que toman valores en el conjunto $\{0,1\}$ y una distribución de probabilidad conjunta dada por $P(A,B,C,D) = P(A)\ P(B)\ P(C \mid A,B)\ P(D \mid C)$. Las distribuciones de probabilidad asociadas son:

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

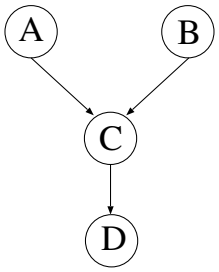
a) Representar gráficamente la red bayesiana

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

c) Dados $B=1, C=1$ y $D=0$, ¿Cuál es la mejor predicción para el valor 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 $P(A \mid B,C,D)$ y calcular su valor para $A=1$ cuando $B=1, C=1$ y $D=0$.

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

$$
P(A=1 \mid B=1,C=1,D=0) = \frac{0.3 \cdot 0.4}{0.7 \cdot 0.2 + 0.3 \cdot 0.4} = 0.4615
$$

c) Dados $B=1, C=1$ y $D=0$, ¿Cuál es el mejor valor de A que se puede predecir?

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

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

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 $A=0$.

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

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