Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 16 de enero de 2015

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

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

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) $\sum_{n=1}^{N} \boldsymbol{x}_n + 1$

B) $\sum_{n=1}^{N} \boldsymbol{x}_n + \lambda\,\boldsymbol{\theta}(k)$

C) $\sum_{n=1}^{N} \boldsymbol{x}_n + \frac{\lambda}{2}$

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

2 $\boxed{\text{C}}$ En el problema de optimización con restricciones

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

se cumplen las condiciones complementarias de Karush-Kuhn-Tucker $\alpha_i^\star v_i(\boldsymbol{\Theta}^\star) = 0$ para $1 \le i \le k$. 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) Para todo i, si $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$,
C) Si para un i, $\alpha_i^\star > 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
D) Existe un i tal que $v_i(\boldsymbol{\Theta}^\star) > 0$ y $\alpha_i^\star = 0$

3 $\boxed{\text{B}}$ 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) El algoritmo EM obtiene los valores óptimos de los parámetros a estimar
C) La verosimilitud del conjunto de entrenamiento, calculada con los parámetros estimados, aumenta en cada iteración del EM
D) En cada iteración, el algoritmo EM estima los valores de las variables ocultas que, en este caso, son los pesos de las gaussianas.

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 : $0 < \alpha_1^* < C$)

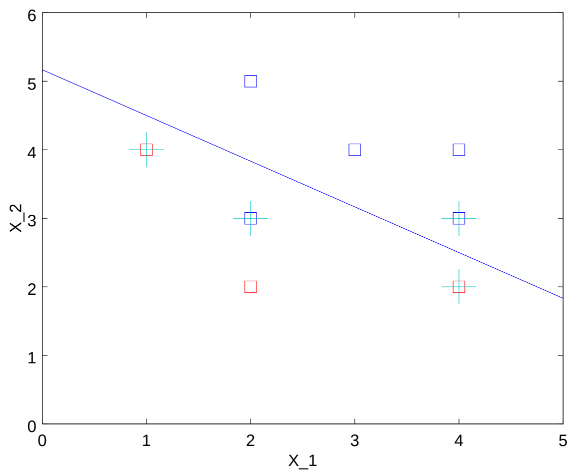
$$\theta_0^\star = c_1 - \boldsymbol{\theta}^{\star t}\mathbf{x_1} = 1 - ((-1.33)\ (1) - (2.00)\ (4)) = 10.33$$

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:

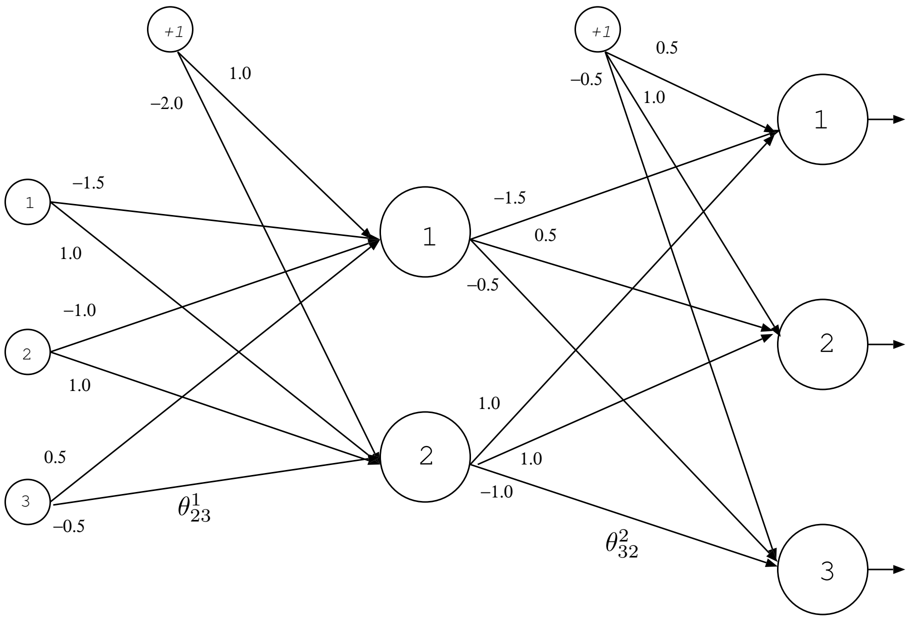


c) Clasificación de la muestra $(5, 5)^t$:

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)

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	$x_1 = 1.0$	$x_2 = 0.0$	$x_3 = 2.0$
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	$t_1 = -1.0$	$t_2 = 1.0$	$t_3 = -2.0$

a) los correspondientes errores 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 1.0

a) Los errores en la capa de salida son:

$\delta^2_1 = (t_1 - s^2_1)\ s^2_1\ (1 - s^2_1) = -0.356$; $\quad \delta^2_2 = (t_2 - s^2_2)\ s^2_2\ (1 - s^2_2) = 0.030$; $\quad \delta^2_3 = (t_3 - s^2_3)\ s^2_3\ (1 - s^2_3) = -0.463$

Los errores en la capa de oculta son:

$\delta^1_1 = (\delta^2_1\ \theta^2_{11} + \delta^2_2\ \theta^2_{21} + \delta^2_3\ \theta^2_{31})\ s^1_1\ (1 - s^1_1) = 0.183$; $\quad \delta^1_2 = (\delta^2_1\ \theta^2_{12} + \delta^2_2\ \theta^2_{22} + \delta^2_3\ \theta^2_{32})\ s^1_2\ (1 - s^1_2) = 0.014$

b) El nuevo peso θ^2_{32} es: $\quad \theta^2_{32} = \theta^2_{32} + \rho\ \delta^2_3\ s^1_2 = (-1.0) + (1)\ (-0.463)\ (0.119) = -1.055$

El nuevo peso θ^1_{23} es: $\quad \theta^1_{23} = \theta^1_{23} + \rho\ \delta^1_2\ x_3 = (-0.5) + (1)\ (0.019)\ (2.0) = -0.472$

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:

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

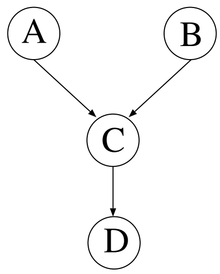
a) Representar gráficamente la red

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 $A=0$.

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

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