Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 3 de febrero de 2021

Apellidos: **Nombre:**   **Grupo:**

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

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

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) Si para un i, $\alpha_i^\star > 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
B) Si para un i, $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
C) Si para un i, $\alpha_i^\star > 0$, entonces $v_i(\boldsymbol{\Theta}^\star) > 0$
D) Si para un i, $\alpha_i^\star = 0$, entonces $u_i(\boldsymbol{\Theta}^\star) = 0$

2 $\boxed{\text{A}}$ En la estimación por máxima verosimilitud de los parámetros de una mezcla de K gaussianas de matriz de covarianza común y conocida a partir de N vectores de entrenamiento, los parámetros a estimar son: el vector-media $\boldsymbol{\mu}_k$ y el peso α_k de cada gaussiana, $k, 1 \leq k \leq K$. Identificar cuál de las siguientes afirnaciones es *correcta*:

A) El método más adecuado es el de *esperanza-maximización* (EM), el cual garantiza que que se cumple la restricción $\sum_{k=1}^K \alpha_k = 1$. Esto es así gracias a que, en cada iteración de EM, los valores de $\alpha_k, 1 \leq k \leq K$, se obtienen como medias de valores de variables latentes, usando una expresión que se deriva analíticamente mediante la técnica de los *multiplicadores de Lagrange* con la restricción indicada.
B) Se puede usar *descenso por gradiente*, ya que los valores de $\boldsymbol{\mu}_k$ no están sujetos a ninguna restricción, lo que hace innecesario recurrir a la técnica de los *multiplicadores de Lagrange*.
C) La solución se obtiene en un paso, utilizando directamente la *optimización lagrangiana* de la verosimilitud de los N vectores de entrenamiento. En este caso, hay un único multiplicador de Lagrange, β, asociado a la restricción de igualdad: $\sum_{k=1}^K \alpha_k = 1$.
D) El método más adecuado sería el de *esperanza-maximización* (EM), pero no es posible utilizarlo ya que EM es un método iterativo que no garantiza el cumplimiento de la restricción de igualdad: $\sum_{k=1}^K \pi_k = 1$.

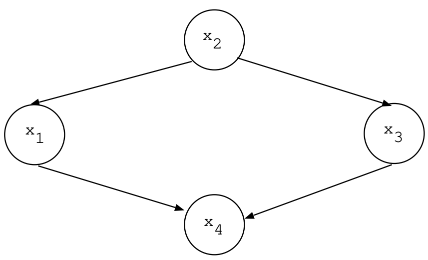
3 $\boxed{\text{D}}$ Se desea ajustar por mínimos cuadrados la función $f : \mathbb{R} \to \mathbb{R}$, definida como: $y = f(x) \stackrel{\text{def}}{=} ax^2 + bx + c$ a una secuencia de N pares entrada-salida: $S = ((x_1, y_1), (x_2, y_2) \ldots, (x_N, y_N))$. La técnica empleada es minimizar por descenso por gradiente la función de error cuadrático:

$$q(a, b, c) = \sum_{n=1}^{N} (f(x_n) - y_n)^2$$

Identifica la afirmación acertada de entre las siguientes:

A) El gradiente es $2ax + b$
B) El descenso por gradiente solo es aplicable a funciones convexas, pero $q(\cdot)$ no lo es.
C) La técnica de descenso por gradiente no es aplicable en este caso ya que la función a ajustar, $f(\cdot)$, no es lineal.
D) El gradiente es: $2\sum_{n=1}^N (f(\boldsymbol{x}_n) - y_n) \; \left[\, x_n^2, x_n, 1\right]^t$

4 $\boxed{\text{A}}$ En la red bayesiana



¿cuál de las relaciones siguientes es falsa en general?

A) $P(x_1, x_3 \mid x_4) = P(x_1 \mid x_4) \, P(x_3 \mid x_4)$
B) $P(x_2, x_4 \mid x_3) = P(x_2 \mid x_3) \, P(x_4 \mid x_3)$
C) $P(x_2, x_4 \mid x_1) = P(x_2 \mid x_1) \, P(x_4 \mid x_1)$
D) $P(x_1, x_3) = P(x_1) \, P(x_3)$

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

En la siguiente tabla se presenta una muestra de entrenamiento no linealmente separable 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
x_{i1}	1	1	1	1	1
x_{i2}	3	4	2	5	1
Clase	-1	$+1$	$+1$	-1	$+1$
$\alpha_i^\star$	10	10	3.56	3.56	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.

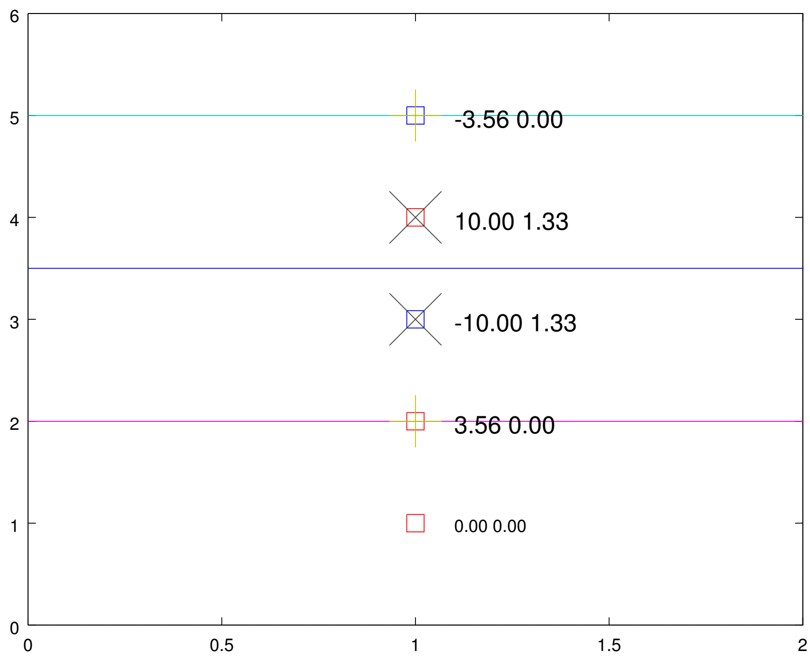
c) Clasificar la muestra $(1, 3.5)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_2\ \alpha_2^\star\ \mathbf{x_2} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_4\ \alpha_4^\star\ \mathbf{x_4}\ \approx\ (0.0, -0.67)$

Usando el vector soporte $\mathbf{x_4}$ (que verifica la condición : $0 < \alpha_4^* < C$) $\quad \theta_0^\star = c_4 - \boldsymbol{\theta}^{\star t}\mathbf{x_4} \approx 2.33$

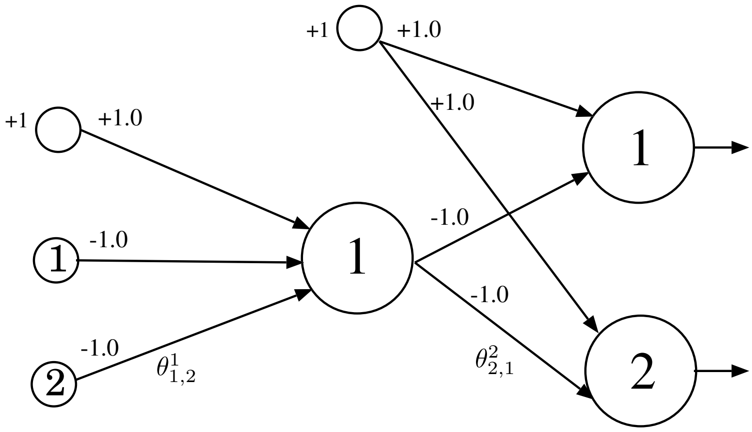
b) Frontera de separación y representación gráfica:
Ecuación de la frontera lineal de separación: $2.33 - 0.67\ x_2\ =\ 0$
y las de los márgenes: $2.33 - 0.67\ x_2\ = +1$ y $2.33 - 0.67\ x_2\ = -1$
Los vectores de entrenamiento son todos los de la tabla. De ellos, los vectores soporte son: $(1,3)^t, (1,4)^t, (1,2)^t, (1,5)^t$.
Representación gráfica:



c) Clasificación de la muestra $(1, 3.5)^t$:
El valor de la función discriminante para este vector es: $2.33 - 0.67\ x_2\ \approx\ -0.015 < 0\ \Rightarrow$ clase -1.

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

El perceptrón multicapa de la figura se utiliza para resolver un problema de regresión, con funciones de activación de los nodos de la capa de salida y el nodo de la capa oculta de tipo *sigmoide*, y factor de aprendizaje $\rho = 1.0$.



Dado un par de entrenamiento $(\boldsymbol{x}^t, t) \; = ((-1,-1),(+1,0))$, calcular:

a) Las salidas de todos los nodos.

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

c) Los nuevos valores de los pesos de las conexiones $\theta^2_{2,1}$ y $\theta^1_{1,2}$.

a) Las salidas de la capa oculta son:

$$\phi^1_1 \;=\; \theta^1_{1,0} + \theta^1_{1,1}\; x_1 + \theta^1_{1,2}\; x_2 \;=\; 3 \qquad s^1_1 \;=\; f_s(\phi^1_1) \;=\; +0.95257$$

Las salida de la capa de salida son:

$$\phi^2_1 \;=\; \theta^2_{1,0} + \theta^2_{1,1}\; s^1_1 \;=\; +0.047426 \qquad s^2_1 \;=\; f_l(\phi^2_1) \;=\; +0.51185$$

$$\phi^2_2 \;=\; \theta^2_{1,0} + \theta^2_{2,1}\; s^1_1 \;=\; +0.047426 \qquad s^2_1 \;=\; f_l(\phi^2_1) \;=\; +0.51185$$

b) Los errores en la capa de salida son:

$$\delta^2_1 \;=\; (t_1 - s^2_1)\; f'_S(\phi^2_1) \;=\; (t_1 - s^2_1)\; s^2_1\; (1 - s^2_1) \;=\; +0.12197$$

$$\delta^2_2 \;=\; (t_2 - s^2_2)\; f'_S(\phi^2_2) \;=\; (t_2 - s^2_2)\; s^2_2\; (1 - s^2_2) \;=\; -0.12789$$

El error en la capa oculta es:

$$\delta^1_1 \;=\; (\delta^2_1\; \theta^2_{1,1} + \delta^2_2\; \theta^2_{1,2})\; f'_S(\phi^1_1) \;=\; (\delta^2_1\; \theta^2_{1,1} + \delta^2_2\; \theta^2_{1,2})\; s^1_1\; (1 - s^1_1) \;=\; 0.0002676$$

c) El nuevo peso $\theta^2_{2,1}$ es:

$$\theta^2_{2,1} \;=\; \theta^2_{2,1} + \Delta\theta^2_{2,1} \;=\; \theta^2_{2,1} + \rho\; \delta^2_2\; s^1_1 \;=\; -1.0 \;+\; 1.0\; (-0.12789)\; 0.95257 \;=\; -1.12183$$

El nuevo peso $\theta^1_{1,2}$ es:

$$\theta^1_{1,2} \;=\; \theta^1_{1,2} + \Delta\theta^1_{1,2} \;=\; \theta^1_{1,2} + \rho\; \delta^1_1\; x_2 \;=\; -1.0 \;+\; 1.0\; (+0.0002676)\; (-1.0) \;=\; -1.0002676$$

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

Las variables aleatorias A, B, C, D, E toman valores en el conjunto $\{0, 1\}$. La distribución de probabilidad conjunta de estas variables viene dada por

$$P(A, B, C, D, E) \;=\; P(A)\ P(B \mid A)\ P(C \mid A, B)\ P(D \mid B, C)\ P(E \mid C, D)$$

con las correspondientes distribuciones de probabilidad:

$$\begin{aligned}
P(A = 1) &\;=\; 0.3 \\
P(B = 1 \mid A = 1) &\;=\; 0.4 \\
P(B = 1 \mid A = 0) &\;=\; 0.6
\end{aligned}$$

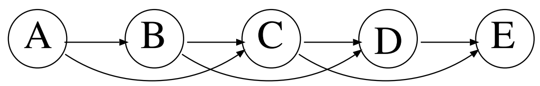
$$\begin{aligned}
P(C = 1 \mid A = 0, B = 0) &= P(D = 1 \mid B = 0, C = 0) = P(E = 1 \mid C = 0, D = 0) = 0.2 \\
P(C = 1 \mid A = 0, B = 1) &= P(D = 1 \mid B = 0, C = 1) = P(E = 1 \mid C = 0, D = 1) = 0.3 \\
P(C = 1 \mid A = 1, B = 0) &= P(D = 1 \mid B = 1, C = 0) = P(E = 1 \mid C = 1, D = 0) = 0.4 \\
P(C = 1 \mid A = 1, B = 1) &= P(D = 1 \mid B = 1, C = 1) = P(E = 1 \mid C = 1, D = 1) = 0.5
\end{aligned}$$

a) Representar gráficamente la red bayesiana correspondiente

b) Obtener una expresión simplificada de $P(D, E \mid A, B, C)$.

c) Dados $A = B = C = 1$, ¿Cuál es la mejor predicción para el valor de E?.

a) Representar gráficamente la red bayesiana correspondiente



b) Obtener una expresión simplificada de $P(D, E \mid A, B, C)$

$$\begin{aligned}
P(D, E \mid A, B, C) &= \frac{P(A, B, C, D, E)}{P(A, B, C)} \\
&= \frac{P(A)\ P(B \mid A)\ P(C \mid A, B)\ P(D \mid B, C)\ P(E \mid C, D)}{\sum_{e,d} P(A)\ P(B \mid A)\ P(C \mid A, B)\ P(D = d \mid B, C)\ P(E = e \mid C, D = d)} \\
&= \frac{\cancel{P(A)}\ \cancel{P(B \mid A)}\ \cancel{P(C \mid A, B)}\ P(D \mid B, C)\ P(E \mid C, D)}{\cancel{P(A)}\ \cancel{P(B \mid A)}\ \cancel{P(C \mid A, B)}\ \sum_{e,d} P(D = d \mid B, C)\ P(E = e \mid C, D = d)} \\
&= \frac{P(D \mid B, C)\ P(E \mid C, D)}{\sum_d P(D = d \mid B, C)\ \sum_e P(E = e \mid C, D = d))} \;=\; P(D \mid B, C)\ P(E \mid C, D)
\end{aligned}$$

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

$$\begin{aligned}
P(D = 0, E = 0 \mid A = 1, B = 1, C = 1) &= P(D = 0 \mid B = 1, C = 1)\ P(E = 0 \mid C = 1, D = 0) = 0.5 \cdot 0.6 = 0.3 \\
P(D = 0, E = 1 \mid A = 1, B = 1, C = 1) &= P(D = 0 \mid B = 1, C = 1)\ P(E = 1 \mid C = 1, D = 0) = 0.5 \cdot 0.4 = 0.2 \\
P(D = 1, E = 0 \mid A = 1, B = 1, C = 1) &= P(D = 1 \mid B = 1, C = 1)\ P(E = 0 \mid C = 1, D = 1) = 0.5 \cdot 0.5 = 0.25 \\
P(D = 1, E = 1 \mid A = 1, B = 1, C = 1) &= P(D = 1 \mid B = 1, C = 1)\ P(E = 1 \mid C = 1, D = 1) = 0.5 \cdot 0.5 = 0.25
\end{aligned}$$

$$P(E \mid A = 1, B = 1, C = 1) \;=\; \sum_d P(D = d, E \mid A = 1, B = 1, C = 1)$$

$$\begin{aligned}
P(E = 0 \mid A = 1, B = 1, C = 1) &= 0.3 + 0.25 = 0.55 \\
P(E = 1 \mid A = 1, B = 1, C = 1) &= 0.2 + 0.25 = 0.45
\end{aligned}$$

$$\hat{e} \;=\; \arg\max_e P(E = e \mid A = 1, B = 1, C = 1) \;=\; 0$$