Examen de los temas 1 a 4 de Aprendizaje Automàtico

ETSINF, Universitat Politècnica de València, 02 de diciembre de 2013

Apellidos: _____ **Nombre:** _____

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas.

1 [B] Deseamos evaluar un sistema de Aprendizaje Automático utilizando un conjunto de datos de entrenamiento que contiene 1000 muestras y la técnica de *exclusión individual* ("Leaving One Out"), obteniéndose un total de 44 errores. Indicar cuál de las afirmaciones siguientes es correcta:

A) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 1000 muestras.
B) La talla de entrenamiento efectiva es de 999 muestras y el error es del 4.4 %
C) La talla de entrenamiento efectiva es de 900 muestras y el error es del 44 %
D) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 900 muestras.

2 [C] Al aplicar la técnica de descenso por gradiente a una modificación de la función de Widrow y Hoff

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

el gradiente $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$ en la iteración $\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) - \rho_k \boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$ es

A) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n$
B) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{\theta}(k)^t\boldsymbol{\theta}(k)$
C) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{\theta}(k)$
D) $\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n + \boldsymbol{x}_n$

3 [C] Entre las siguientes propiedades de las funciones discriminantes lineales hay una que es falsa:

A) La función discriminante lineal aplicada en un punto devuelve un valor proporcional a la distancia del punto al correspondiente hiperplano separador.
B) La distancia del origen de coordenadas al hiperplano separador asociado a una función discriminante lineal es $\frac{\theta_0}{\|\boldsymbol{\theta}\|}$
C) Un hiperplano separador tiene asociado una única función discriminante lineal canónica
D) Un hiperplano separador tiene asociado un número infinito de funciones discriminantes lineales

4 [B] Se quiere aplicar la técnica esperanza-maximización a un problema de estimación de máxima verosimilitud en el que no hay variables latentes o ocultas. En este caso ¿Cuál de las afirmaciones siguientes es correcta?

A) En ese caso no se puede aplicar la técnica esperanza-maximización.
B) En ese caso solo se aplica la etapa de maximización y en una iteración acaba.
C) En ese caso solo se aplica la etapa de maximización y hay que iterar hasta que converja.
D) En ese caso solo se aplica la etapa del cálculo de la esperanza.

Problema 1 (4 puntos; tiempo estimado: 40 minutos)

En una tarea de clasificación de correos electrónicos como spam o no-spam se dispone de un conjunto S de 500 correos *no-spam* (clase A) y 300 *spam* (clase B).

a) ¿Cuál sería el logaritmo de la verosimilitud de los mensajes de S si las probabilidades a priori $P(A)$ y $P(B)$ fueran iguales?

b) Las probabilidades a priori puede estimarse por máxima verosimilitud a partir de S como: $P(A) = 5/8,\ P(B) = 3/8$. Derivar estas probabilidades mediante la técnica de optimización de los multimplicadores de Lagrange

c) Calcular el logaritmo de la verosimilitud de S según las probabilidades a prioiri obtenidas en b) y compararla con la obtenida en a)

a)

- Modelo: $\boldsymbol{\Theta}_0 \equiv (p_A, p_B)^t$: $p_A \equiv P(c = A) = 0.5,\ p_B \equiv P(c = B) = 0.5$
- El logaritmo (neperiano) de la verosimilitud es

$$\log P(S \mid \boldsymbol{\Theta}_0) \ = \ \log(\prod_{i=1}^{500} p_A \prod_{j=1}^{300} p_B) \ = \ 500 \cdot \log 0.5 + 300 \cdot \log 0.5 \ = -554.52 \ \Rightarrow \ P(S \mid \boldsymbol{\Theta}_0) \ = \ 1.496 \cdot 10^{-241}$$

b)

- Modelo: $\boldsymbol{\Theta} \equiv (p_A, p_B)^t$, con $p_A + p_B = 1$
- Verosimilitud y logaritmo de la verosimilitud:

$$\begin{aligned} P(S \mid \boldsymbol{\Theta}) \ &= \ \prod_{i=1}^{500} p_A \prod_{j=1}^{300} p_B \ = \ p_A^{500} p_B^{300} \\ L_S(\boldsymbol{\Theta}) \ &= \ \log P(S \mid \boldsymbol{\Theta}) \ = \ 500 \log p_A + 300 \log p_B \end{aligned}$$

- Estimación de máxima verosimilitud:

$$\boldsymbol{\Theta}^\star \ = \ \arg\max_{\boldsymbol{\Theta}} L_S(\boldsymbol{\Theta}) \ = \ \arg\max_{\substack{p_A, p_B \\ p_A + p_B = 1}} \ (500 \log p_A + 300 \log p_B)$$

- Lagrangiana: $\Lambda(p_A, p_B, \beta) = 500 \log p_A + 300 \log p_B + \beta\,(1 - p_A - p_B)$
- Soluciones óptimas en función del multiplicador de Lagrange:

$$\left.\begin{aligned} \frac{\partial \Lambda}{\partial p_A} = \frac{500}{p_A} - \beta = 0 \\ \frac{\partial \Lambda}{\partial p_B} = \frac{300}{p_B} - \beta = 0 \end{aligned}\right\} \quad \begin{aligned} p_A^\star(\beta) = \frac{500}{\beta} \\ p_B^\star(\beta) = \frac{300}{\beta} \end{aligned}$$

- Función dual de Lagrange:

$$\Lambda_D(\beta) \ = \ 500 \log \frac{500}{\beta} + 300 \log \frac{300}{\beta} + \beta(1 - \frac{500}{\beta} - \frac{300}{\beta}) \ = \ \beta - 800 \log \beta - 800 + 500 \log 500 + 300 \log 300$$

- Valor óptimo del multiplicador de Lagrange: $\dfrac{d\Lambda_D}{d\beta} \ = \ 1 - \dfrac{800}{\beta} \ = \ 0 \ \Rightarrow \ \beta^\star \ = \ 800$
- Solución final: $\boldsymbol{\theta}^\star \ = \ (p_A^\star, p_B^\star)^t$: $\quad p_A^\star = p_A^\star(\beta^\star) = \dfrac{5}{8} \quad p_B^\star = p_B^\star(\beta^\star) = \dfrac{3}{8}$

c) Como en a):

$$L_S(\boldsymbol{\Theta}^\star) \ = \ 500 \cdot \log(5/8) + 300 \cdot \log(3/8) \ = -529.25 \ \Rightarrow \ P(S \mid \boldsymbol{\Theta}^\star) = 1.411 \cdot 10^{-230} \ \gg \ 1.496 \cdot 10^{-241} = P(S \mid \boldsymbol{\Theta}_0)$$

La verosmilitud es mayor que en a) debido a que se ha maximizado la verosimilitud con respecto a $\boldsymbol{\Theta}$.

Problema 2 (4 puntos; tiempo estimado: 20 minutos)

Para el aprendizaje de una máquina de vectores soporte se dispone de una muestra de entrenamiento linealmente separable

$$S = \{((1,4),+1),((1,6),+1),((2,2),+1),((2,3),+1),((4,2),-1),((3,4),-1),((3,5),-1),((5,4),-1),((5,6),-1)\}$$

Los multiplicadores de Lagrange óptimos son: $\boldsymbol{\alpha}^* = (0,\ 0.25,\ 0,\ 1.0,\ 0,\ 1.25,\ 0,\ 0,\ 0)^t$.

- Obtener la función discriminante lineal correspondiente
- Calcular el margen óptimo
- Clasificar la muestra $(4,5)$.

- La función discriminante lineal
  - El vector de pesos:

$$\theta_1^\star = +1\cdot 0.25\cdot 1 \ +1\cdot 1.0\cdot 2 \ -1\cdot 1.25\cdot 3 = -1.5$$
$$\theta_2^\star = +1\cdot 0.25\cdot 6 \ +1\cdot 1.0\cdot 3 \ -1\cdot 1.25\cdot 4 = -0.5$$

  - El peso umbral (con la muestra 2) $\theta_0^* = (+1) - (-1.5\ 1 - 0.5\ 6) = 5.5$
  - La FDL: $\phi(\boldsymbol{x}) = -1.5\ x_1 - 0.5\ x_2 + 5.5$
- Margen óptimo:

$$\frac{2}{\|\boldsymbol{\theta}^\star\|} = \frac{2}{\sqrt{0.25+1.0+0,1.25}} = 1.26$$

- Clasificación de la muestra $(4,5)$: $\phi(4,5) = -1.5\cdot 4 - 0.5\cdot 5 + 5.5 = -3 < 0 \Rightarrow$ clase $= -1$

Examen de los temas 5 a 7 de Aprendizaje Automàtico

ETSINF, Universitat Politècnica de València, 16 de enero de 2014

Apellidos: _____ **Nombre:** _____

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas.

1 [A] Si la función de acción no lineal en la capa de salida de un perceptrón de dos capas fuese lineal, las fórmulas que permiten modificar los pesos de dicha capa de salida en el algoritmo BackProp verifican que (solo una respuesta es correcta):

A) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ s^1_j$
B) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ g(\phi^2_i)\ (1 - g(\phi^2_i))\ s^1_j$
C) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ g(\phi^2_i)\ s^1_j$
D) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ (1 - g(\phi^2_i))\ s^1_j$

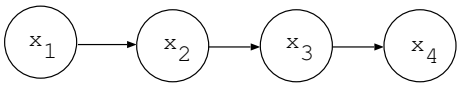
2 [B] En una presentación de las N muestras de aprendizaje mediante el algoritmo de retropropagación del error, indicar qué afirmación es la correcta:

A) Los pesos se modifican una sola vez tanto en la versión "online" o incremental como en la versión batch.
B) Los pesos se modifican N veces en la versión "online" o incremental y una en la versión batch.
C) Los pesos se modifican N veces en la versión "online" y también en la versión batch.
D) Los pesos no se modifican en la versión "online" o incremental y una en la versión batch.

3 [B] En el perceptrón multicapa, la parálisis de la red se produce cuando (solo una respuesta es correcta)

A) En test, los pesos son muy grandes
B) En entrenamiento, los valores de las combinaciones lineales de los nodos son muy grandes
C) En test, los valores de las combinaciones lineales de los nodos son muy grandes
D) En entrenamiento, los pesos son nulos

4 [A] En la red bayesiana lineal

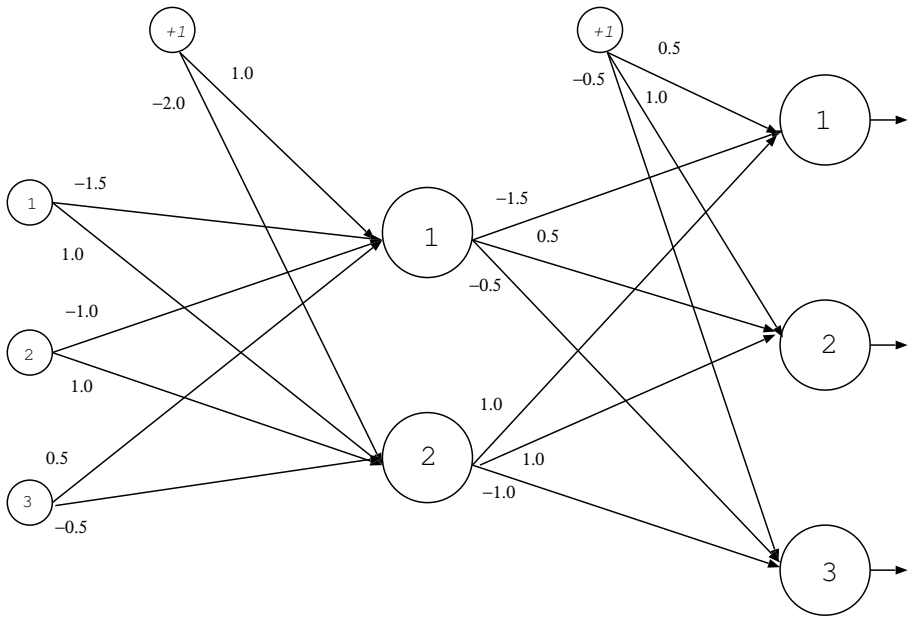


¿cuál de las relaciones siguientes es correcta?

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

Problema 1 (4 puntos; tiempo estimado: 30 minutos)

En el perceptrón multicapa de la figura con funciones sigmoid en todos los nodos



se aplica el algoritmo de retropropagación del error para la muestra de entrenamiento $(\boldsymbol{x}, \boldsymbol{t})$ con $\boldsymbol{x} = (1, 0, -2)$ y $\boldsymbol{t} = (0, 1, 0)$ con un factor de aprendizaje $\rho = 1.0$ y en la fase ascendente se obtienen las siguientes salidas (las datos numéricos están redondeados a la primera cifra decimal para facilitar el cálculo)

- Capa oculta: $\phi_1^1 = -1.5 \quad s_1^1 = 0.2$
 $\phi_2^1 = 0.0 \quad s_2^1 = 0.5$
- Capa de salida: $\phi_1^2 = 0.2 \quad s_1^2 = 0.5$
 $\phi_2^2 = 0.6 \quad s_2^2 = 0.7$
 $\phi_3^2 = -0.6 \quad s_3^2 = 0.3$

a) Calcular el error que se observan en el nodo 2 de la capa oculta

b) Calcular los pesos finales del nodo 1 al nodo 2 de la capa de salida y del nodo 1 al nodo 2 de la capa de oculta después de una iteración

c) Si las funciones de activación de la capa de salida fuesen lineales y las salidas de la capa de salida para $\boldsymbol{x}$ fueran $s_1^2 = 0.6 \;\; s_2^2 = 0.6 \;\; s_3^2 = 0.4$ y las salidas de la capa de oculta fueran $s_1^1 = 0.2 \;\; s_2^1 = 0.5$, calcular los pesos finales del nodo 1 al nodo 2 de la capa de salida después de una iteración

a) Errores en la capa salida:

$\delta_1^2 = (t_1 - s_1^2)\; g'(\phi_1^2) = (t_1 - s_1^2)\; s_1^2\; (1 - s_1^2) = -0.125$

$\delta_2^2 = (t_2 - s_2^2)\; g'(\phi_2^2) = (t_2 - s_2^2)\; s_2^2\; (1 - s_2^2) = 0.063$

$\delta_3^2 = (t_3 - s_3^2)\; g'(\phi_3^2) = (t_3 - s_3^2)\; s_3^2\; (1 - s_3^2) = -0.063$

Errores en el nodo 2 de la capa oculta:

$\delta_2^1 = \left(\delta_1^2\; \theta_{12}^2 + \delta_2^2\; \theta_{22}^2 + \delta_3^2\; \theta_{32}^2\right)\; g'(\phi_2^1) = \left(\delta_1^2\; \theta_{12}^2 + \delta_2^2\; \theta_{22}^2 + \delta_3^2\; \theta_{32}^2\right)\; s_2^1\; (1 - s_2^1) = 0.00025$

b) Modificaciones de los pesos:

Peso del nodo 1 de la capa oculta al nodo 2 de la capa de salida:

$\Delta\theta_{21}^2 = \rho\; \delta_2^2\; s_1^1 = 0.0126 \Rightarrow \theta_{21}^2 = 0.5 + 0.0126 = 0.5126$

Peso del nodo 1 de la capa de entrada al nodo 2 de la capa oculta:

$\Delta\theta_{21}^1 = \rho\; \delta_2^1\; x_1 = 0.00025 \Rightarrow \theta_{21}^1 = 1.0 + 0.00025 = 1.00025$

c) Caso de funciones de activación lineal en la capa de salida:

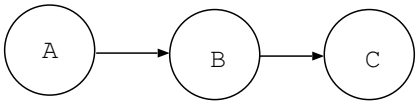
$\delta_2^2 = (t_2 - s_2^2) = 1 - 0.6 = 0.4$

Peso del nodo 1 de la capa oculta al nodo 2 de la capa de salida:

$\Delta\theta_{21}^2 = \rho\; \delta_2^2\; s_1^1 = 0.08 \Rightarrow \theta_{21}^2 = 0.5 + 0.08 = 0.58$

Problema 2 (4 puntos; tiempo estimado: 30 minutos)

En la red bayesiana de la figura, las variables A, B y C toman valores en el conjunto $\{0, 1\}$ y las distribuciones de probabilidad asociadas son:



$$P(A = 1) = 0.3$$
$$P(B = 1 \mid A = 1) = 0.2 \quad P(B = 1 \mid A = 0) = 0.9$$
$$P(C = 1 \mid B = 1) = 0.1 \quad P(C = 1 \mid B = 0) = 0.4$$

a) Calcular $P(B = 0 \mid A = 1, C = 0)$

b) Calcular $P(A = 1 \mid B = 1, C = 1)$

$$P(A, B, C) = P(A)\ P(B \mid A)\ P(C \mid B)$$

a) Cálculo de $P(B = 0 \mid A = 1, C = 0)$:

$$\begin{aligned} P(B = 0 \mid A = 1, C = 0) &= \frac{P(A = 1, B = 0, C = 0)}{P(A = 1, C = 0)} \\ &= \frac{P(A = 1)P(B = 0 \mid A = 1)P(C = 0 \mid B = 0)}{P(A = 1)\sum_{b \in 0,1} P(B = b \mid A = 1)P(C = 0 \mid B = b)} \\ &= \frac{0.3 \cdot 0.8 \cdot 0.6}{0.3(0.8 \cdot 0.6 \ + \ 0.2 \cdot 0.9)} = \frac{0.144}{0.198} = 0.727 \end{aligned}$$

b) Cálculo de $P(A = 1 \mid B = 1, C = 1)$. Como A y C son independientes cuando se conoce B:

$$\begin{aligned} P(A = 1 \mid B = 1, C = 1) &= P(A = 1 \mid B = 1) \\ &= \frac{P(A = 1, B = 1)}{P(B = 1)} \\ &= \frac{\sum_{c \in \{0,1\}} P(A = 1, B = 1, C = c)}{\sum_{a,c \in \{0,1\}} P(A = a, B = 1, C = c)} \\ &= \frac{P(A = 1)\ P(B = 1 \mid A = 1)\ \sum_{c \in \{0,1\}}\ P(C = c \mid B = 1)}{\sum_{a \in \{0,1\}} P(A = a)\ P(B = 1 \mid A = a)\ \sum_{c \in \{0,1\}} P(C = c \mid B = 1))} \\ &= \frac{P(A = 1)\ P(B = 1 \mid A = 1)}{\sum_{a \in \{0,1\}} P(A = a)\ P(B = 1 \mid A = a)} \\ &= \frac{0.3\ 0.2}{0.3\ 0.2 + 0.7\ 0.9} = \frac{0.06}{0.69} = 0.086957 \end{aligned}$$

Examen de recuperación de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 28 de enero de 2014

Apellidos: _____ **Nombre:** _____

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas.

1 [B] Se ha evaluado un sistema de Aprendizaje Automático mediante la técnica de *validación cruzada en B bloques* ("B-fold Cross Validation") con $B = 10$ y utilizando un conjunto de datos de entrenamiento que contiene 1000 muestras. Se han obtenido un total de 20 errores. Indicar cuál de las afirmaciones siguientes es correcta:

A) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 1000 muestras.
B) La talla de entrenamiento efectiva es de 900 muestras y el error es del 2 %
C) La talla de entrenamiento efectiva es de 900 muestras y el error es del 20 %
D) La talla de entrenamiento efectiva es de 1000 muestras y y el error es del 20 %.

2 [B] Considerar la siguiente modificación de la función de Widrow y Hoff

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

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) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n$
B) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \boldsymbol{\theta}(k)$
C) $\displaystyle\sum_{n=1}^{N} \left(\boldsymbol{\theta}(k)^t \boldsymbol{x}_n - y_n\right) \boldsymbol{x}_n + \boldsymbol{\theta}(k)^t\boldsymbol{\theta}(k)$
D) $\displaystyle\sum_{n=1}^{N} \boldsymbol{\theta}(k)^t \boldsymbol{x}_n + 1$

3 [A] Un clasificador implementado mediante una red neuronal con L capas ocultas en la que todas las funciones de activación son lineales, es equivalente a:

A) un clasificador basado en funciones discriminantes lineales
B) un clasificador basado en funciones discriminantes lineales generalizadas cuyas fronteras de decisión son no-lineales
C) un clasificador implementado mediante una red neuronal con $L - 1$ capas ocultas
D) para que la red pueda usarse para clasificación, al menos las funciones de activación de la capa de salida han de ser no-lineales.

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

A) todas las probabilidades condicionales e incondicionales en las que participan VA's de $\mathcal{C}$ se pueden obtener mediante las reglas basicas de inferencia estadística, a partir de la probabilidad conjunta de todas las VA de $\mathcal{C}$.
B) la probabilidad incondicional de una VA $a \in \mathcal{C}$ solo depende de las VA's de los nodos con los que está conectado el nodo de a.
C) el grafo que representa las VA's de $\mathcal{C}$ ha de ser acíclico y conexo.
D) ha de existir indepenencia condicional entre al menos dos VA's de $\mathcal{C}$.

Problema 1 (4 puntos; tiempo estimado: 30 minutos)

En la siguiente tabla se muestra una muestra de entrenamiento linealmente separable en $\mathbb{R}^2$ y los correspondientes multiplicadores de Lagrange óptimos obtenidos al entrenar una máquina de vectores soporte con esta muestra:

i	1	2	3	4	5	6	7	8	9
x_{i1}	1	1	2	2	4	3	3	5	5
x_{i2}	2	3	2	1	1	6	3	1	2
Clase	+1	+1	+1	+1	-1	-1	-1	-1	-1
$\alpha_i^\star$	0	0	1.111	0	0.222	0	0.889	0	0

a) Obtener la función discriminante lineal correspondiente

b) Dibujar la función discriminante lineal

c) Clasificar la muestra $(5,5)^t$.

a) Pesos de la función discriminante:

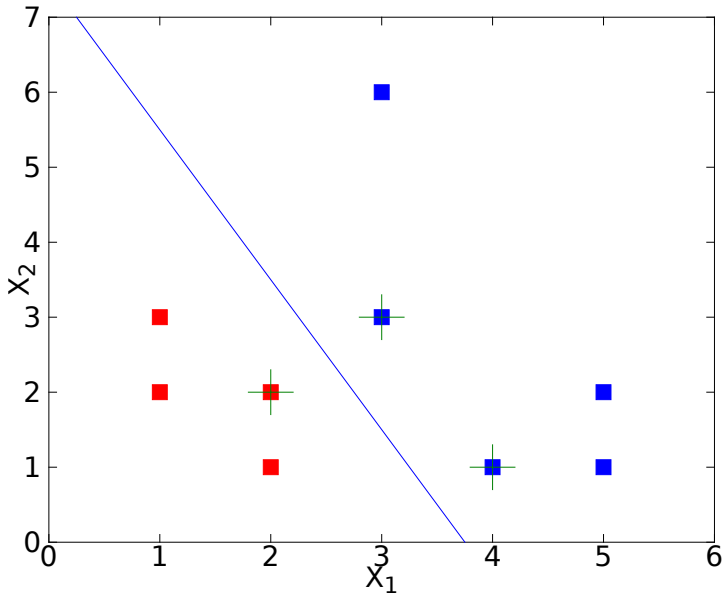
$\boldsymbol{\theta}^\star = c_3\alpha_3^\star\mathbf{x_3} + c_5\alpha_5^\star\mathbf{x_5} + c_7\alpha_7^\star\mathbf{x_7}$

$\theta_1^* = (+1)\ (2)\ (1.111) + (-1)\ (4)\ (0.222) + (-1)\ (3)\ (0.889) = -1.333$
$\theta_2^* = (+1)\ (2)\ (1.111) + (-1)\ (1)\ (0.222) + (-1)\ (3)\ (0.889) = -0.667$

Usando el vector soporte $\mathbf{x_3}$:

$\theta_0^\star = c_3 - \boldsymbol{\theta}^{\star t}\mathbf{x_3} = 1 - (-1.333 \cdot 2 - 0.667 \cdot 2) = 5.000$

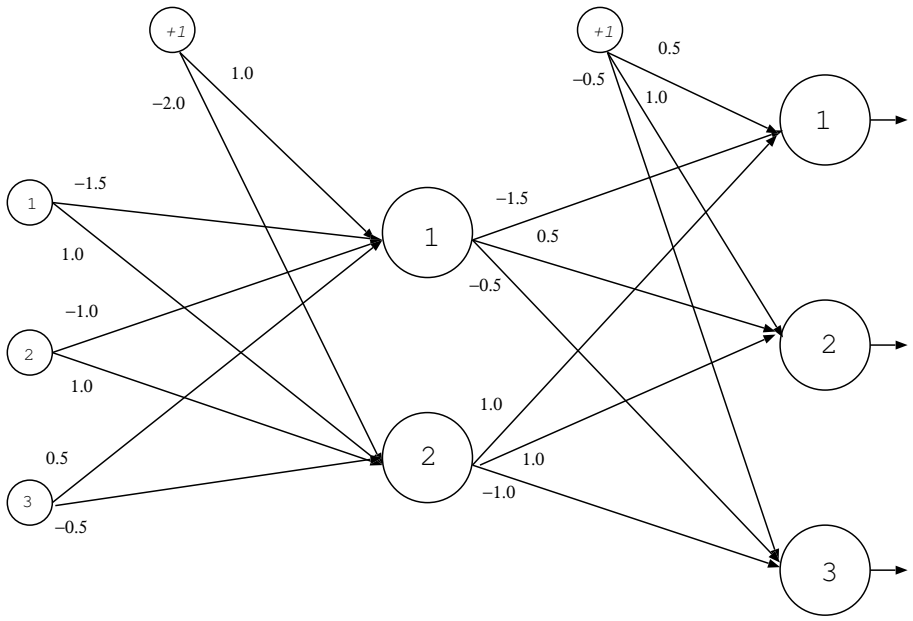
b) Dibujo de la función discriminante:



c) Clasificación de la muestra $(5,5)^t$:
$\theta_0^* + \theta_1^*\ 5 + \theta_2^*\ 5 = -5 < 0 \Rightarrow$ clase -1 (o la 2)

Problema 2 (4 puntos; tiempo estimado: 30 minutos)

Sea $\mathcal{P}$ el perceptrón multicapa de la figura, que se pretende utilizar para resolver un problema de regresión.



Se asume que la función de activación de los nodos de la capa de salida es lineal y la de los nodos de la capa oculta es de tipo escalón, definida como:

$$g_E(z) = \text{sgn}(z) = \begin{cases} -1 & \text{if } z < 0 \\ +1 & \text{if } z \geq 0 \end{cases}$$

Sea $\mathbf{x} = (2.0,\ 2.0,\ 0.0)^t$ un vector de entrada y sean $-5.0,\ 1.0, -2.0,$ los valores deseados para dicha muestra en los nodos de salida $1, 2, 3$, respectivamente. Calcular:

a) los tres valores de salida de $\mathcal{P}$ cuando se observa $\mathbf{x}$ en la entrada,

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

a) Valores de la capa oculta:

$s_1^1 = -1; s_2^1 = 1$

Valores de la capa de salida:

$s_1^2 = 3; s_2^2 = 1.5; s_3^2 = -1.0$

b) Los errores en la capa de salida son:

$\delta_1^2 = t_1 - s_1^1 = -8.0; \quad \delta_2^2 = t_2 - s_2^1 = -0.5; \quad \delta_3^2 = t_3 - s_3^1 = -1.0$

La función en escalón no tiene derivada en 0, y vale 0 en cualquier otro punto, por lo tanto los errores en la capa oculta no se pueden calcular cuando la función discriminante vale 0. Como las funciones discriminates en la capa oculta para el vector son -4 y 2, el error en la capa oculta sería 0. No obstante, la propagación del error de la capa de salida a la oculta sin contar la derivada de la función de activación es:

nodo 1: $\delta_1^2\ \theta_{11}^2 + \delta_2^2\ \theta_{21}^2 + \delta_3^2\ \theta_{31}^2 = 12.25; \quad$ nodo 2: $\delta_1^2\ \theta_{12}^2 + \delta_2^2\ \theta_{22}^2 + \delta_3^2\ \theta_{32}^2 = -7.5$

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 [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: $\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) - \rho_k \nabla q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$. En esta expresión, el gradiente, $\nabla q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$, 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 [C] 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) 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 [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 [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 : $0 < \alpha_1^* < C$)

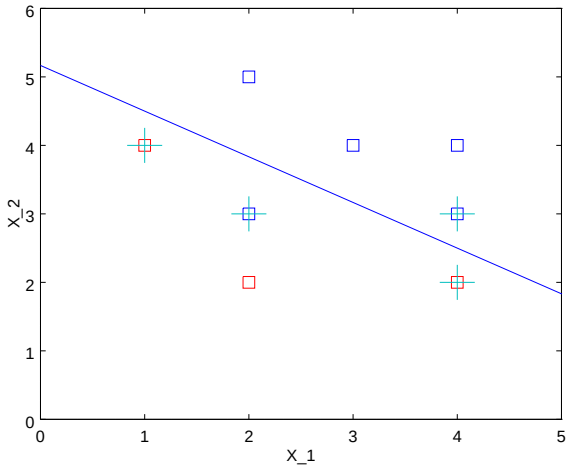
$\theta_0^\star = c_1 - \boldsymbol{\theta}^{\star t}\mathbf{x_1} = 1 - ((-1.33)\ (1) - (2.00)\ (4)) = 10.33$

b) Frontera de separación y representación gráfica:

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

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:

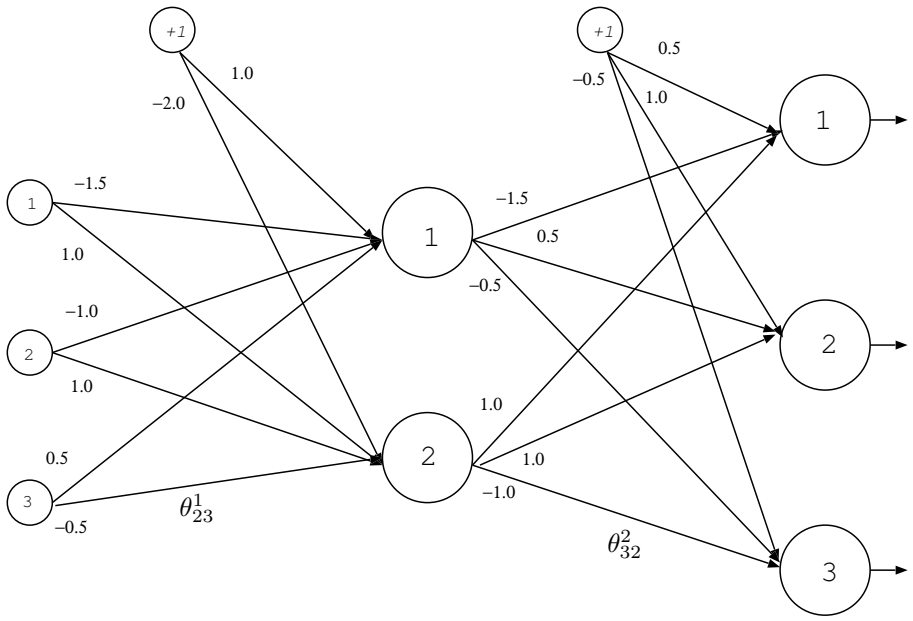


c) Clasificación de la muestra $(5,5)^t$:

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:

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

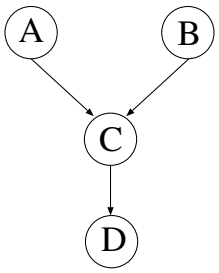
a) Representar gráficamente la red

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 $A=0$.

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

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

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 28 de enero de 2015

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 [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 de entrenamiento 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 $2.2\,\% \pm 0.7\,\%$
D) El error estimado es $22\,\% \pm 7\,\%$.

2 [A] Considerar la siguiente modificación de la función de Widrow y Hoff

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

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

3 [A] Considerar el aprendizaje mediante máquinas de vectores soportes y márgenes blandos con una muestra de aprendizaje $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ no separable linealmente. Si un multiplicador de Lagrange óptimo α_j^*, asociado a la restricción c_j ($\boldsymbol{\theta}^t \boldsymbol{x}_d jn + \theta_0$) $\geq 1 - \zeta_j$, $1 \leq j \leq N$, es cero, entonces:

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.

4 [A] La distribución de probabilidad conjunta en una red bayesiana de tres nodos A, B y C es $P(A, B, C) = P(C)\; P(A \mid C)\; P(B \mid C)$. Marcar cuál es la afirmación correcta:

A) $P(A, B \mid C) = P(A \mid C)\; P(B \mid C)$
B) En general, $P(A, B \mid C) \neq P(A \mid C)\; P(B \mid C)$
C) $P(A, B \mid C) = P(C \mid A)\; P(C \mid B)$
D) $P(A, B \mid C) = P(A)\; P(B)$

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	3	2	4	3	2	4	4
x_{i2}	4	1	3	2	4	5	4	3
Clase	+1	+1	+1	+1	−1	−1	+1	−1
$\alpha_i^\star$	3.38	0	0	5.75	9.13	0	10	10

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 $(1,1)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_4\ \alpha_4^\star\ \mathbf{x_4} + c_5\ \alpha_5^\star\ \mathbf{x_5} + c_7\ \alpha_7^\star\ \mathbf{x_7} + c_8\ \alpha_8^\star\ \mathbf{x_8}$

$\theta_1^* = (+1)\ (1)\ (3.38) + (+1)\ (4)\ (5.75) + (-1)\ (3)\ (9.13) + (+1)\ (4)\ (10) + (-1)\ (4)\ (10) \approx -1.0$

$\theta_2^* = (+1)\ (4)\ (3.38) + (+1)\ (2)\ (5.75) + (-1)\ (4)\ (9.13) + (+1)\ (4)\ (10) + (-1)\ (3)\ (10) = -1.5$

Usando el vector soporte $\mathbf{x_4}$ (que verifica la condición : $0 < \alpha_4^* < C$)

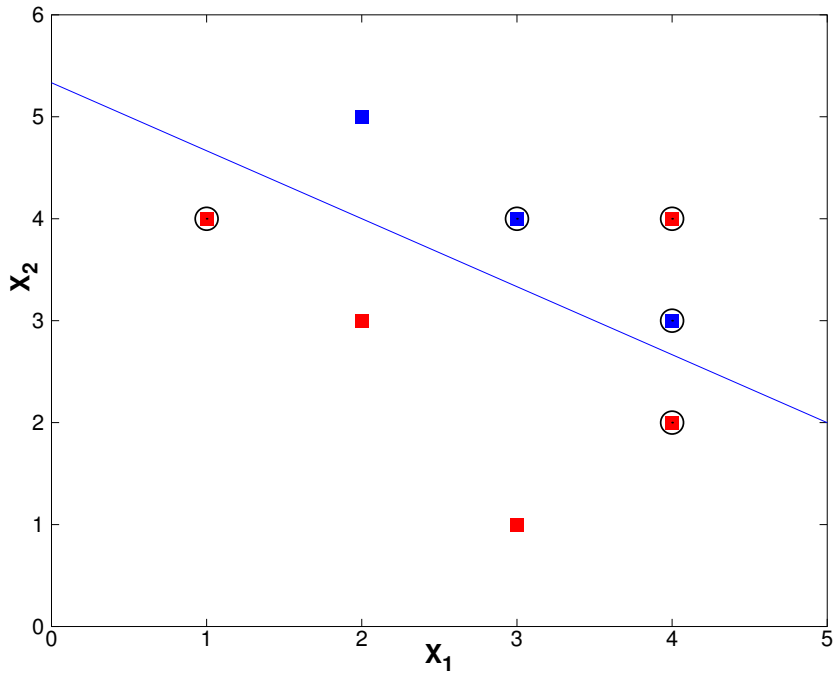
$\theta_0^\star = c_4 - \boldsymbol{\theta}^{\star t}\mathbf{x_4} = 1 - ((-1.0)\ (4) + (-1.5)\ (2)) = 8.0$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $8.0 - 1.0\ x_1 - 1.5\ x_2 = 0 \rightarrow x_2 \approx -0.67\ x_1 + 5.3$.

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

Representación gráfica:

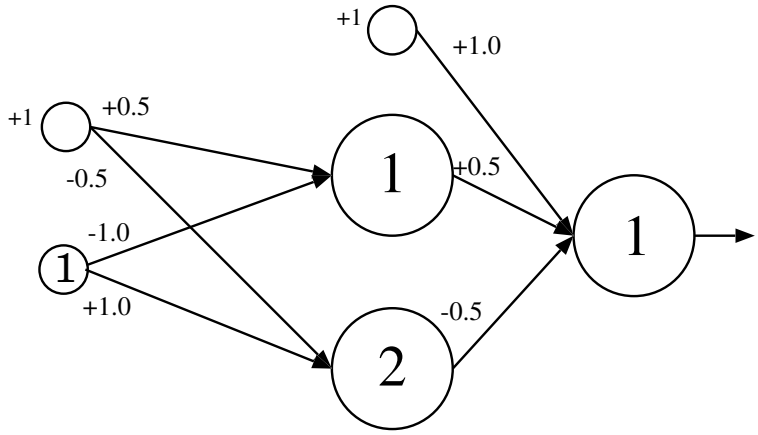


c) Clasificación de la muestra $(1,1)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 1 + \theta_2^*\ 1 = 5.5 > 0 \Rightarrow$ 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 función de activación de los nodos de la capa de salida y de la capa oculta de tipo sigmoid, y factor de aprendizaje $\rho = 1.0$.



Dado un vector de entrada $x = 1$ y su valor deseado de salida $t = -1$, Calcular:

a) las salidas de todas las unidades

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

c) Los nuevos valores de los pesos de las conexiones θ^1_{21}, que va del nodo 1 de entrada al nodo 2 de la capa oculta, y θ^2_{12}, que va del nodo 2 de la capa oculta al nodo de la capa de salida.

a) Las salidas de todas las unidades

$$\begin{aligned} \phi^1_1 = \theta^1_{11}\, x_1 + \theta^1_{10} &= -0.5\,; & s^1_1 = \tfrac{1}{1+\exp(-\phi^1_1)} &= 0.378 \\ \phi^1_2 = \theta^1_{21}\, x_1 + \theta^1_{10} &= 0.5\,; & s^1_2 = \tfrac{1}{1+\exp(-\phi^1_2)} &= 0.622 \\ \phi^2_1 = \theta^2_{11}\, s^1_1 + \theta^2_{12}\, s^1_2 + \theta^2_{10} &= 0.878\,; & s^2_1 = \tfrac{1}{1+\exp(-\phi^2_1)} &= 0.706 \end{aligned}$$

a) El error en la capa de salida es: $\delta^2_1 = (t_1 - s^2_1)\; s^2_1\; (1 - s^2_1) = -0.354$
Los errores en la capa de oculta son: $\delta^1_1 = \delta^2_1\; \theta^2_{11}\; s^1_1\; (1 - s^1_1) = -0.042$; $\delta^1_2 = \delta^2_1\; \theta^2_{12}\; s^1_2\; (1 - s^1_2) = 0.042$

b) El nuevo peso θ^2_{12} es: $\theta^2_{12} = \theta^2_{12} + \rho\; \delta^2_1\; s^1_2 = (-0.5) + (1)\; (-0.354)\; (0.622) = -0.720$
El nuevo peso θ^1_{21} es: $\theta^1_{21} = \theta^1_{21} + \rho\; \delta^1_2\; x_1 = (+1.0) + (1)\; (0.042)\; (1.0) = 1.042$

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

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

- $P(R)$ es uniforme: $P(R = 1) = P(R = 2) = P(R = 3)$
- $P(X \mid R)$, $P(Y \mid R)$ y $P(Z \mid R)$ son idénticas y vienen dadas en la tabla T.

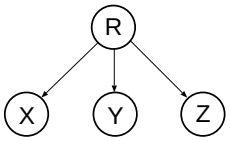
T	`"a"`	`"b"`	`"c"`
1	1/3	0	2/3
2	1/4	1/2	1/4
3	0	3/5	2/5

a) Representar gráficamente la red

b) Obtener una expresión simplificada de $P(X, Y, Z \mid R)$ en función de las distribuciones que definen $\mathcal{R}$ y calcular $P(X =$ `"a"`$, Y =$ `"a"`$, Z =$ `"a"`$ \mid R = 1)$

c) Calcular $P(R = 3 \mid X =$ `"b"`$, Y =$ `"b"`$, Z =$ `"b"`$)$

a) Representación gráfica de la red:



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

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

$$P(X =\texttt{"a"}, Y =\texttt{"a"}, Z =\texttt{"a"} \mid R = 1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot = \frac{1}{27}$$

c)

$$P(R = 3 \mid X =\texttt{"b"}, Y =\texttt{"b"}, Z =\texttt{"b"}) = \frac{P(R = 3, X = "b", Y = "b", Z = "b")}{P(X = "b", Y = "b", Z = "b")}$$

$$= \frac{P(R = 3)\ P(X = "b" \mid R = 3)\ P(Y = "b" \mid R = 3)\ P(Z = "b" \mid R = 3)}{\sum_{r \in \{1,2,3\}} P(R = r)\ P(X = "b" \mid R = r)\ P(Y = "b" \mid R = r)\ P(Z = "b" \mid R = r)}$$

$$= \frac{\frac{1}{3}\ \frac{3}{5}\ \frac{3}{5}\ \frac{3}{5}}{\frac{1}{3}\ 0\ 0\ 0\ +\ \frac{1}{3}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ +\ \frac{1}{3}\ \frac{3}{5}\ \frac{3}{5}\ \frac{3}{5}} = \frac{216}{341} \approx 0.633$$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 11 de enero de 2016

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 [D] Se ha evaluado un sistema de Aprendizaje Automático mediante la técnica de *Exclusión individual* ("Leaving One Out") y utilizando un conjunto de datos que contiene 200 muestras. Se han obtenido un total de 10 errores. Indicar cuál de las afirmaciones siguientes es razonable:

A) La talla de entrenamiento efectiva es 190 muestras, la del test es de 10 muestras y el error estimado es $5.0\,\% \pm 0.3\,\%$
B) La talla de entrenamiento efectiva es de 199 muestras, la del test es de 1 muestra y el error estimado es $5.0 \pm 3.0\,\%$
C) La talla de entrenamiento efectiva es de 200 muestras, la del test es de 10 muestras y el error estimado es $5.0 \pm 0.3\,\%$
D) La talla de entrenamiento efectiva es de 199 muestras, la del test es de 200 muestras y el error estimado es $5.0 \pm 3.0\,\%$

2 [D] En el problema de optimización con restricciones

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

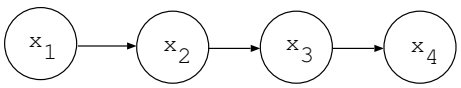
se cumplen las condiciones complementarias de Karush-Kuhn-Tucker $\alpha_i^\star v_i(\boldsymbol{\Theta}^\star) = 0$ para $1 \leq i \leq k$. Indicar cuál de las siguientes afirmaciones se deduce de ellas:

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

3 [C] 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 π_k de cada gaussiana, $k, 1 \leq k \leq K$. Identificar cuál de las siguientes afirnaciones es *correcta*:

A) 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*.
B) 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: $\sum_{k=1}^K \pi_k = 1$.
C) El método más adecuado es el de *esperanza-maximización* (EM), el cual garantiza que que se cumple la restricción $\sum_{k=1}^K \pi_k = 1$. Esto es así gracias a que, en cada iteración de EM, los valores de $\pi_k, 1 \leq k \leq K$, 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.
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: $\sum_{k=1}^K \pi_k = 1$.

4 [B] En la red bayesiana lineal



¿cuál de las relaciones siguientes es falsa en general?

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

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}	1	2	3	4	5
Clase	$+1$	$+1$	-1	$+1$	-1
$\alpha_i^\star$	0	3.56	10	10	3.56

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 $(1, 4.5)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_2\ \alpha_1^\star\ \mathbf{x_2} + c_3\ \alpha_4^\star\ \mathbf{x_3} + c_4\ \alpha_5^\star\ \mathbf{x_4} + c_5\ \alpha_7^\star\ \mathbf{x_5}$

$\theta_1^* = 0.0$

$\theta_2^* \approx -0.67$

Usando el vector soporte $\mathbf{x_5}$ (que verifica la condición : $0 < \alpha_5^* < C$)

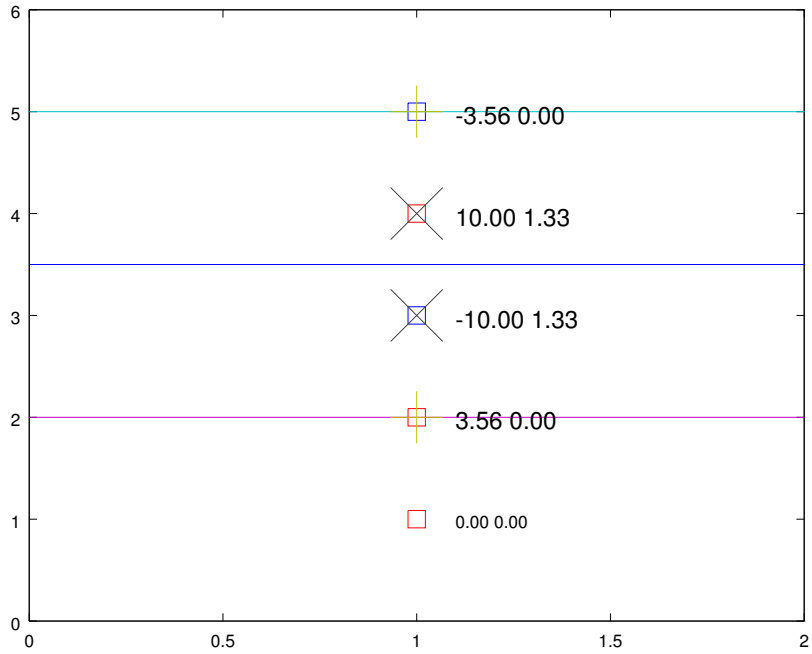
$\theta_0^\star = c_5 - \boldsymbol{\theta}^{\star t}\mathbf{x_5} \approx 2.33$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $2.33 - 0.67\ x_2 = 0$

Los vectores de entrenamiento son todos los de la tabla. De ellos, los vectores soporte son: $(1,2)^t, (1,3)^t, (1,4)^t, (1,5)^t$.

Representación gráfica:

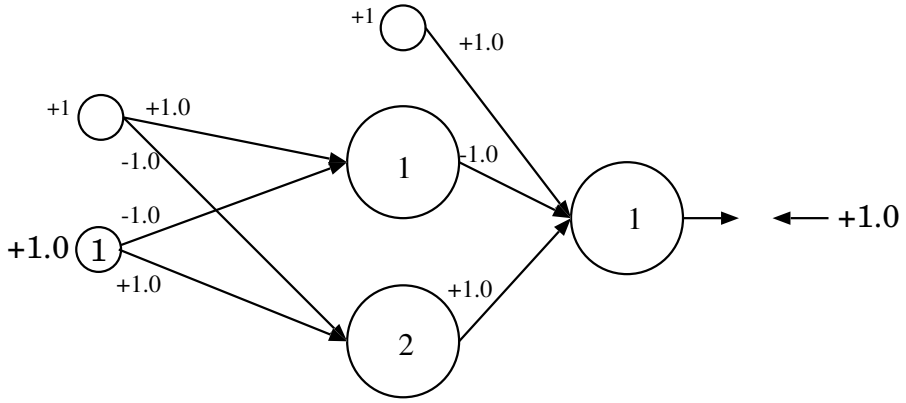


c) Clasificación de la muestra $(1, 4.5)^t$:

El valor de la función discriminante para este vector es: $2.33 - 0.67\ x_2 \approx -0.67 < 0 \Rightarrow$ 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 función de activación de los nodos de la capa de salida y de la capa oculta de tipo *tangente hiperbólica*, y factor de aprendizaje $\rho = 0.5$.



Dado un vector de entrada $x = 1$ y su valor deseado de salida $t = +1$, Calcular:

a) las salidas de todas las unidades

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

c) Los nuevos valores de los pesos de las conexiones θ^1_{21}, que va del nodo 1 de entrada al nodo 2 de la capa oculta, y θ^2_{12}, que va del nodo 2 de la capa oculta al nodo de la capa de salida.

a) Las salidas de todas las unidades

$$\phi^1_1 = \theta^1_{11}\ x_1 + \theta^1_{10}\ = 0.0\,; \qquad s^1_1 = \frac{\exp(\phi^1_1)-\exp(-\phi^1_1)}{\exp(\phi^1_1)+\exp(-\phi^1_1)}\ =\ 0.0$$

$$\phi^1_2 = \theta^1_{21}\ x_1 + \theta^1_{10}\ =\ \ 0.0\,; \qquad s^1_2 = \frac{\exp(\phi^1_2)-\exp(-\phi^1_2)}{\exp(\phi^1_2)+\exp(-\phi^1_2)}\ =\ 0.0$$

$$\phi^2_1 = \theta^2_{11}\ s^1_1 + \theta^2_{12}\ s^1_2 + \theta^2_{10}\ =\ \ 1.0\,; \qquad s^2_1 = \frac{\exp(\phi^2_1)-\exp(-\phi^2_1)}{\exp(\phi^2_1)+\exp(-\phi^2_1)}\ =\ 0.76159$$

b) El error en la capa de salida es:

$\delta^2_1 = (t_1 - s^2_1)\ (1 - (s^2_1)^2) = +0.10012$

Los errores en la capa de oculta son:

$\delta^1_1 = \delta^2_1\ \theta^2_{11}\ (1 - (s^1_1)^2) = -0.10012; \quad \delta^1_2 = \delta^2_1\ \theta^2_{12}\ (1 - (s^1_2)^2) = +0.10012$

c) El nuevo peso θ^2_{12} es: $\theta^2_{12}\ =\ \theta^2_{12} + \rho\ \delta^2_1\ s^1_2\ =\ (+1.0) + (0.5)\ (+0.10012)\ (0.0)\ =\ 1.0$

El nuevo peso θ^1_{21} es: $\theta^1_{21}\ =\ \theta^1_{21} + \rho\ \delta^1_2\ x_1\ =\ (+1.0) + (0.5)\ (+0.10012)\ (1.0)\ =\ 1.0501$

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

Considerar la red bayesiana $\mathcal{R}$ definida como $P(R, X, Y, Z) = P(R)\ P(X \mid R)\ P(Y \mid R)\ P(Z \mid R)$, 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:

- $P(R)$ es uniforme: $P(R = 1) = P(R = 2) = P(R = 3)$
- $P(X \mid R)$, $P(Y \mid R)$ y $P(Z \mid R)$ son idénticas y vienen dadas en la tabla T.

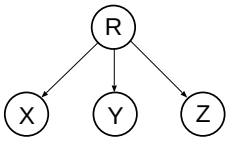
T	"a"	"b"	"c"
1	1/3	0	2/3
2	1/4	1/2	1/4
3	0	3/5	2/5

a) Representar gráficamente la red

b) Obtener una expresión simplificada de $P(X, Y, Z \mid R)$ en función de las distribuciones que definen $\mathcal{R}$ y calcular $P(X = \texttt{"a"}, Y = \texttt{"a"}, Z = \texttt{"a"} \mid R = 1)$

c) Calcular $P(R = 3 \mid X = \texttt{"b"}, Y = \texttt{"b"}, Z = \texttt{"b"})$

a) Representación gráfica de la red:



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

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

$$P(X = \texttt{"a"}, Y = \texttt{"a"}, Z = \texttt{"a"} \mid R = 1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot = \frac{1}{27}$$

c)

$$P(R = 3 \mid X = \texttt{"b"}, Y = \texttt{"b"}, Z = \texttt{"b"}) = \frac{P(R = 3, X = \text{"}b\text{"}, Y = \text{"}b\text{"}, Z = \text{"}b\text{"})}{P(X = \text{"}b\text{"}, Y = \text{"}b\text{"}, Z = \text{"}b\text{"})}$$

$$= \frac{P(R = 3)\ P(X = \text{"}b\text{"} \mid R = 3)\ P(Y = \text{"}b\text{"} \mid R = 3)\ P(Z = \text{"}b\text{"} \mid R = 3)}{\sum_{r \in \{1,2,3\}} P(R = r)\ P(X = \text{"}b\text{"} \mid R = r)\ P(Y = \text{"}b\text{"} \mid R = r)\ P(Z = \text{"}b\text{"} \mid R = r)}$$

$$= \frac{\frac{1}{3}\ \frac{3}{5}\ \frac{3}{5}\ \frac{3}{5}}{\frac{1}{3}\ 0\ 0\ 0\ +\ \frac{1}{3}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ +\ \frac{1}{3}\ \frac{3}{5}\ \frac{3}{5}\ \frac{3}{5}} \approx 0.633$$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 27 de enero de 2016

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 [A] Identifica cuál de las siguientes afirmaciones es *errónea o impropia*:

A) La *teoría de la decisión estadística* es idónea para problemas de *clasificación*, pero es inadecuada para problemas de *regresión* en aprendizaje de funciones.
B) *Clasificación* puede considerarse como un caso particular de *regresión* y los problemas de *sobreajuste* y *sobregeneralización* le afectan de forma similar
C) En el *aprendizaje activo* se asume que un agente externo al sistema se encarga de etiquetar datos de entrenamiento seleccionados por el sistema
D) El *aprendizaje adaptativo* es esencialmente un modo *supervisado* de aprendizaje

2 [A] Se ha evaluado un sistema de Aprendizaje Automático mediante un proceso de *validación cruzada en B bloques* ("B-fold Cross Validation") con $B = 8$ y 1000 muestras etiquetadas. En este proceso se han producido 15 errores en total. Indicar cuál de las afirmaciones siguientes es razonable:

A) La talla de entrenamiento efectiva es 875 muestras y el error estimado es $1.5\,\% \pm 0.75\,\%$
B) La talla de entrenamiento efectiva es de 992 muestras y el error estimado es inferior al $2\,\%$
C) La talla de test efectiva es de 125 muestras y el error estimado es $12\,\% \pm 5.7\,\%$.
D) Como en cada bloque de test solo hay 125 muestras, el error es muy variable, pudiéndose estimar como $1.5\,\% \pm 5.7\,\%$.

3 [A] Se desea ajustar por mínimos cuadrados la función $f : \mathbb{R}^2 \to R$, definida como: $y = f(\boldsymbol{x}) \stackrel{\text{def}}{=} ax_1^2 + bx_2^2 + cx_1x_2$ a una secuencia de N pares entrada-salida: $S = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots, (\boldsymbol{x}_N, y_N))$. La técnica empleada es minimizar por descenso por gradiente la función de error cuadrático:

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

Identifica la afirmación acertada de entre las siguientes:

A) El vector gradiente es: $2\sum_{n=1}^{N}(f(\boldsymbol{x}_n) - y_n) \cdot \left(x_{n1}^2,\ x_{n2}^2,\ x_{n1}x_{n2} \right)^t$
B) El gradiente es $2ax_1 + 2bx_2 + cx_1x_2$
C) El descenso por gradiente solo es aplicable a funciones convexas, pero $q(\cdot)$ no lo es.
D) La técnica de descenso por gradiente no es aplicable en este caso ya que la función a ajustar, $f(\cdot)$, no es lineal.

4 [A] Considerar el aprendizaje mediante máquinas de vectores soportes y márgenes blandos con una muestra de aprendizaje $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ no separable linealmente. Si un multiplicador de Lagrange óptimo α_j^*, asociado a la restricción $c_j\ (\boldsymbol{\theta}^t \boldsymbol{x}_d jn + \theta_0)\ \geq\ 1 - \zeta_j$, $1 \leq j \leq N$, es cero, entonces:

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.

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
x_{i1}	4	1	2	3	4	2
x_{i2}	1	2	2	2	2	3
Clase	-1	$+1$	-1	$+1$	-1	$+1$
$\alpha_i^\star$	0	3.11	10.00	10.00	3.78	0.67

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 $(1,1)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_2\ \alpha_2^\star\ \mathbf{x_2} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_4\ \alpha_4^\star\ \mathbf{x_4} + c_5\ \alpha_5^\star\ \mathbf{x_5} + c_6\ \alpha_6^\star\ \mathbf{x_6}$

$\theta_1^* = (+1)\ (3.11)\ (1) + (-1)\ (10)\ (2) + (+1)\ (10)\ (3) + (-1)\ (3.78)\ (4) + (+1)\ (0.67)\ (2) \approx -0.67$

$\theta_2^* = (+1)\ (3.11)\ (2) + (-1)\ (10)\ (2) + (+1)\ (10)\ (2) + (-1)\ (3.78)\ (2) + (+1)\ (0.67)\ (3) \approx +0.67$

Usando el vector soporte $\mathbf{x_5}$ (que verifica la condición : $0 < \alpha_5^* < C = 10$)

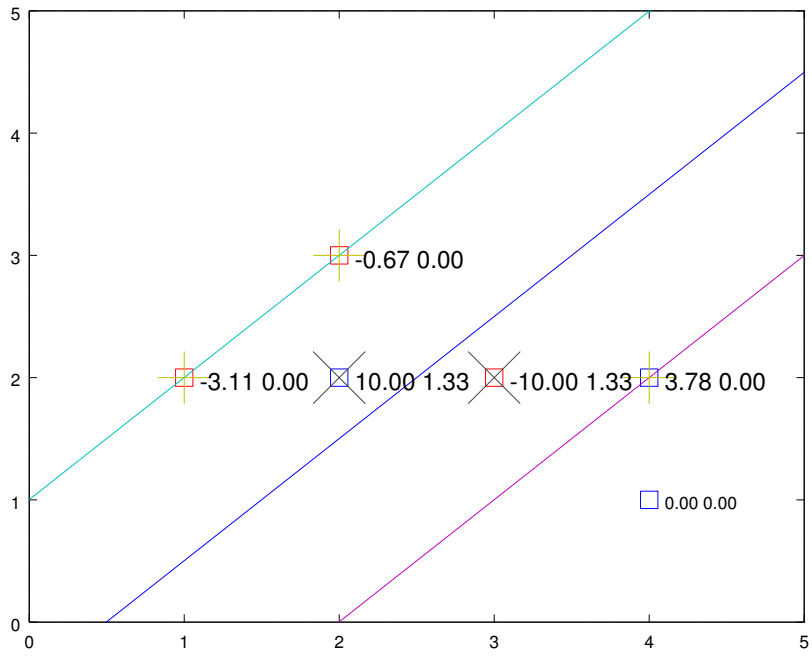
$\theta_0^\star = c_5 - \boldsymbol{\theta}^{\star t}\mathbf{x_5} = -1 - ((-0.67)\ (4) + (0.67)\ (2)) = 0.34$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $-0.67\ x_1 + 0.67\ x_2 + 0.34 = 0$

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

Representación gráfica:

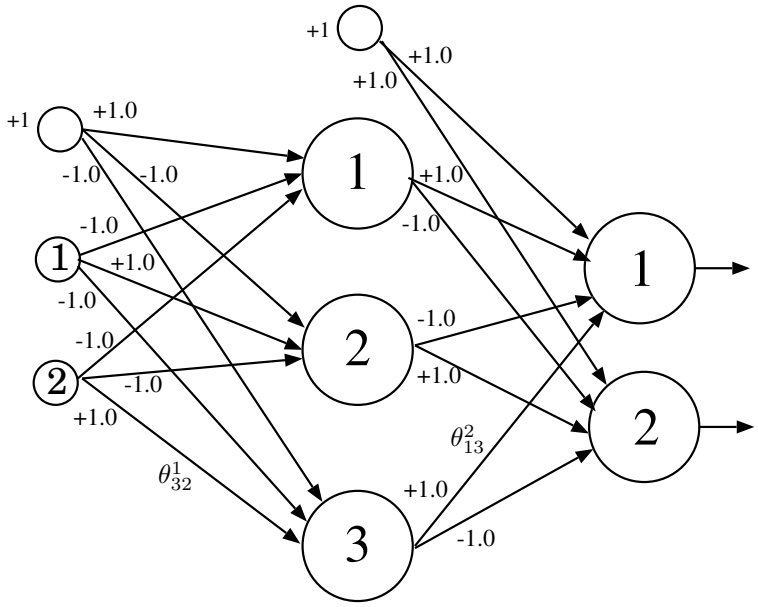


c) Clasificación de la muestra $(1,1)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 1 + \theta_2^*\ 1 = +0.34 > 0 \Rightarrow$ 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 función de activación de los nodos de la capa de salida y de la capa oculta de tipo *sigmoide*, y factor de aprendizaje $\rho = 1.0$.



Dado un vector de entrada $\boldsymbol{x}^t = (1, -1)$, las salidas de las unidades de la capa de salida son $s_1^2 = 0.740$ y $s_2^2 = 0.722$ y las de las unidades ocultas son $s_1^1 = 0.731$, $s_2^1 = 0.731$ y $s_3^1 = 0.047$. Si el valor deseado de salida es $\boldsymbol{t}^t = (1, 0)$, calcular:

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

b) Los nuevos valores de los pesos de las conexiones θ^1_{32} y θ^2_{13}.

a) Los errores en la capa de salida son:

$\delta_1^2 = (t_1 - s_1^2)\, s_1^2\, (1 - s_1^2) = +0.050;$

$\delta_2^2 = (t_2 - s_2^2)\, s_2^2\, (1 - s_2^2) = -0.145$

Los errores en la capa de oculta son:

$\delta_1^1 = (\delta_1^2\, \theta_{11}^2 + \delta_2^2\, \theta_{21}^2)\, s_1^1\, (1 - s_1^1) = +0.038;$

$\delta_2^1 = (\delta_1^2\, \theta_{12}^2 + \delta_2^2\, \theta_{22}^2)\, s_2^1\, (1 - s_2^1) = -0.038;$

$\delta_3^1 = (\delta_1^2\, \theta_{13}^2 + \delta_2^2\, \theta_{23}^2)\, s_3^1\, (1 - s_3^1) = +0.009$

b) El nuevo peso θ_{13}^2 es: $\theta_{13}^2 = \theta_{13}^2 + \rho\, \delta_1^2\, s_3^1 = (+1.0) + (1)\, (0.050)\, (0.047) = 1.002;$

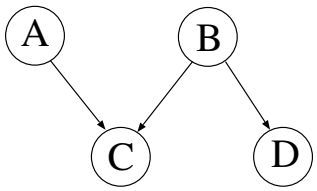
El nuevo peso θ_{32}^1 es: $\theta_{32}^1 = \theta_{32}^1 + \rho\, \delta_3^1\, x_2 = (+1.0) + (1)\, (0.009)\, (-1.0) = 0.991$

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

Las variables aleatorias A, B, C, D toman valores en el conjunto $\{0, 1\}$. La distribución de probabilidad conjunta de estas variables viene dada por $P(A, B, C, D) = P(A)\ P(B)\ P(C \mid A, B)\ P(D \mid B)$, donde las distribuciones de probabilidad asociadas son:

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

Y cuya representación gráfica es



a) Calcular la probabilidad conjunta para $A = 1,\ B = 1, C = 1$ y $D = 1$.

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 $D = 1$.

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

a) Calcular la probabilidad conjunta para $A = 1,\ B = 1, C = 1$ y $D = 1$.

$$P(A=1, B=1, C=1, D=1) = P(A=1)\ P(B=1)\ P(C \mid A=1, B=1)\ P(D \mid B=1) = 0.3\,0.4\,0.4\,0.7 = 0.0336$$

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 $D = 1$.

$$
\begin{aligned}
P(A \mid B, C, D) &= \frac{P(A, B, C, D)}{P(B, C, D)} = \frac{P(A)\ P(B)\ P(C \mid A, B)\ P(D \mid B)}{P(B)\ P(D \mid B)\ \sum_a P(A=a)\ P(C \mid A=a, B)} \\
&= \frac{P(A)\ P(C \mid A, B)}{P(A=0)\ P(C \mid A=0, B) + P(A=1)\ P(C \mid A=1, B)}
\end{aligned}
$$

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

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

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

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

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 11 de enero de 2017

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 [C] Sea S un conjunto de datos supervisados o etiquetados. Para el diseño de un sistema de reconocimiento de formas, se utilizan datos de S tanto para aprender los parámetros del modelo de reconocimiento, $\mathcal{M}$, como para estimar la probabilidad de error de reconocimiento esperada para dicho modelo, p_e. Indicar cual de las siguientes afirmaciones es incorrecta.

A) Si S es suficientemente grande, el método de *validación cruzada en B bloques* puede proporcionar buenas estimaciones de p_e, basadas en todos los datos de S. Una vez estimado p_e, también es recomendable usar todos los datos de S para el aprendizaje final de $\mathcal{M}$.

B) Si la talla de S es 160, y se desea que el intervalo de confianza al 95 % de p_e sea menor que $\pm 1\,\%$, el método de *partición* sería totalmente inapropiado.

C) Si S es suficientemente grande, se puede elegir un valor adecuado de B para que el método de *validación cruzada en B bloques* garantice un entrenamiento de $\mathcal{M}$ que evite tanto el sobreajuste como el sobreenternamiento.

D) Si se usa el método de *exclusión individual* con un conjunto S cuya talla es menor de 100 y se obtiene $p_e = 0.1$, el intervalo de confianza al 95 % de esta estimación será mayor que $\pm 5\,\%$.

2 [D] En el problema de optimización con restricciones

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

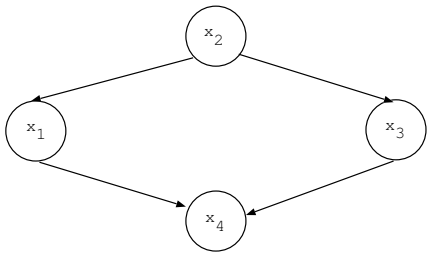
se cumplen las condiciones complementarias de Karush-Kuhn-Tucker $\alpha_i^\star v_i(\boldsymbol{\Theta}^\star) = 0$ para $1 \leq i \leq k$. Indicar cuál de las siguientes afirmaciones se deduce de ellas:

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

3 [C] Las siguientes afirmaciones se refieren al método Esperanza Maximización (EM) aplicado a una muestra de entrenamiento S. Identificar cuál de ellas es errónea o inapropiada:

A) EM es útil para estimar valores maximo-verosímiles de los parámetros de modelos estadísticos a partir de S.
B) EM es un método iterativo que garantiza la convergencia a un máximo local de la verosimilitud de S.
C) La rapidez de convergencia de EM puede mejorarse eligiendo un factor de aprendizaje adecuado para S.
D) La rapidez de convergencia de EM puede mejorarse inicializando los parámetros de forma adecuada para S.

4 [D] En la red bayesiana



¿cuál de las relaciones siguientes es falsa en general?

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

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

Para el aprendizaje de una máquina de vectores soporte se dispone de la siguiente muestra de entrenamiento linealmente separable:

$S = \{((1,1),+1),((1,4),+1),((1,6),+1),((2,2),+1),((2,3),+1),((4,2),-1),((3,4),-1),((3,5),-1),((5,5),-1),((6,4),-1)\}$

Los multiplicadores de Lagrange óptimos son: $\boldsymbol{\alpha}^* = (0, 0, 0.25, 0, 1.0, 0, 1.25, 0, 0, 0)^t$.

a) Obtener la función discriminante lineal correspondiente

b) Obtener la ecuación de la frontera de decisión entre clases y representar gráficamente los puntos de entrenamiento y dicha frontera.

c) Calcular el margen óptimo

d) Clasificar la muestra $(3, 1)$.

a) Función discriminante lineal (FDL):

- Vector de pesos:

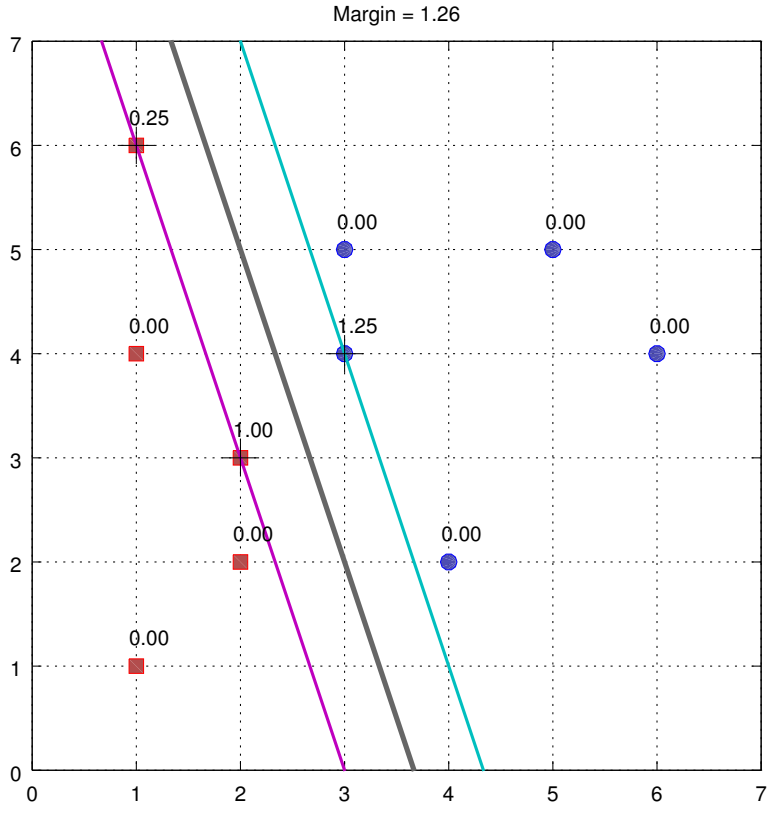
$$\theta_1^\star = +1 \cdot 0.25 \cdot 1 \ +1 \cdot 1.0 \cdot 2 \ -1 \cdot 1.25 \cdot 3 = -1.5$$
$$\theta_2^\star = +1 \cdot 0.25 \cdot 6 \ +1 \cdot 1.0 \cdot 3 \ -1 \cdot 1.25 \cdot 4 = -0.5$$

- Peso umbral (con el tercer vector de entrenamiento): $\theta_0^* = (+1) - (-1.5 \cdot 1 - 0.5 \cdot 6) = 5.5$
- FDL: $\phi(\boldsymbol{x}) = -1.5\ x_1 - 0.5\ x_2 + 5.5$

b) Ecuación de la frontera de decisión:

$$-1.5\ x_1 - 0.5\ x_2 + 5.5 = 0 \quad \Rightarrow \quad x_2 = -3x_1 + 11$$

Representación gráfica:



c) Margen óptimo:

$$\frac{2}{\|\boldsymbol{\theta}^\star\|} = \frac{2}{\sqrt{(-1.5)^2 + (-0.5)^2}} = 1.265$$

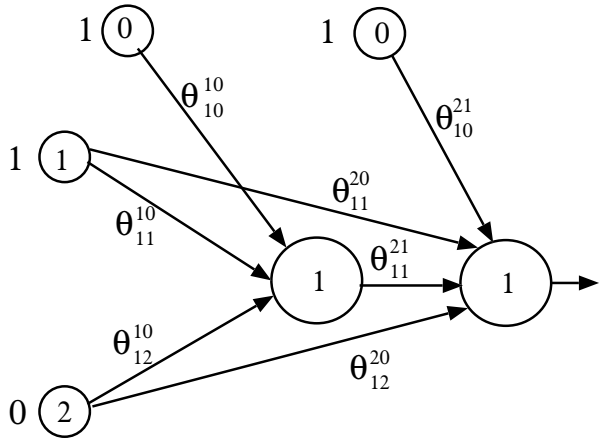
Alternativamente:

$$2\left(\sum_{n\in\mathcal{V}} \alpha_n^\star\right)^{-1/2} = \frac{2}{\sqrt{0.25 + 1.0 + 1.25}} = 1.265$$

d) Clasificación de la muestra $(3, 1)$: $\phi(3,1) = -1.5 \cdot 3 - 0.5 \cdot 1 + 5.5 = 0.5 > 0 \quad \Rightarrow \quad \text{clase} = +1$

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

La red hacia adelante ("feedforward") de la figura se utiliza para resolver un problema de regresión, con función de activación de los nodos de la capa de salida y de la capa oculta de tipo *sigmoid*, y factor de aprendizaje $\rho = 1.0$.



Dados unos pesos iniciales $\theta_{10}^{10} = \theta_{11}^{10} = \theta_{12}^{10} = \theta_{11}^{20} = \theta_{12}^{20} = \theta_{10}^{21} = \theta_{11}^{21} = 1.0$, un vector de entrada $\boldsymbol{x}^t = (1, 0)$ y su valor deseado de salida $t = +1$, Calcular:

a) las salidas de todas las unidades

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

c) Los nuevos valores de los pesos de las conexiones

Pista: La actualización de pesos en esta red sigue la misma formulación que en el BackProp para el perceptrón multicapa convencional: el incremento de peso es $\Delta\theta = \rho\ z\ \delta$, donde ρ es el factor de aprendizaje, z es la entrada del arco asociado al peso θ, y δ es el error que se observa en la salida de la unidad a la que llega ese arco, multiplicado por la derivada de la función de activación.

a) Las salidas de todas las unidades

$\phi_1^1 = \theta_{10}^{10} + \theta_{11}^{10}\ x_1 + \theta_{12}^{10}\ x_2 = 2.0$ $\qquad$ $s_1^1 = \dfrac{1}{1+\exp(-\phi_1^1)} = .880797$

$\phi_1^2 = \theta_{10}^{21} + \theta_{11}^{20}\ x_1 + \theta_{10}^{20}\ x_2 + \theta_{11}^{21}\ s_1^1 = 2.880797$ $\qquad$ $s_1^2 = \dfrac{1}{1+\exp(-\phi_1^2)} = .946889$

b) El error en la capa de salida es:

$\delta_1^2 = (t_1 - s_1^2)\ s_1^2\ (1 - s_1^2) = .002671$

El error en la capa de oculta es:

$\delta_1^1 = (\delta_1^2\ \theta_{11}^{21})\ s_1^1\ (1 - s_1^1) = .000280$

c) Los nuevos pesos son:

$\theta_{10}^{21} = \theta_{10}^{21} + \rho\ \delta_1^2\ (+1) = \theta_{10}^{21} + 0.002671 = 1.002671$

$\theta_{11}^{21} = \theta_{11}^{21} + \rho\ \delta_1^2\ s_1^2 = \theta_{11}^{21} + 0.002529 = 1.002529$

$\theta_{11}^{20} = \theta_{11}^{20} + \rho\ \delta_1^2\ x_1 = \theta_{11}^{20} + 0.002671 = 1.002671$

$\theta_{12}^{20} = \theta_{12}^{20} + \rho\ \delta_1^2\ x_2 = \theta_{12}^{20} + 0.0 = 1.0$

$\theta_{10}^{10} = \theta_{10}^{10} + \rho\ \delta_1^1\ (+1) = \theta_{10}^{10} + 0.000280 = 1.000280$

$\theta_{11}^{10} = \theta_{11}^{10} + \rho\ \delta_1^1\ x_1 = \theta_{11}^{10} + 0.000280 = 1.000280$

$\theta_{12}^{10} = \theta_{12}^{10} + \rho\ \delta_1^1\ x_2 = \theta_{11}^{10} + 0.0 = 1.0$

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

Considerar la red bayesiana $\mathcal{R}$ definida como $P(W,X,Y,Z) = P(W)\ P(X \mid W)\ P(Y \mid X)\ P(Z \mid X)$, cuyas variables aleatorias, W, X, Y, Z, toman valores en en el conjunto {a,b,c}. Las distribuciones de probabilidad asociadas son como sigue:

- $P(W)$ es uniforme: $P(W=\text{a}) = P(W=\text{b}) = P(W=\text{c})$,
- $P(X \mid W)$, $P(Y \mid X)$ y $P(Z \mid X)$ vienen dadas en las siguientes tablas:

$P(x \mid w)$	x: a	b	c
w: a	1/2	0	1/2
b	1/4	1/2	1/4
c	1/5	3/5	1/5

$P(y \mid x)$	y: a	b	c
x: a	1/3	0	2/3
b	1/4	1/2	1/4
c	0	3/5	2/5

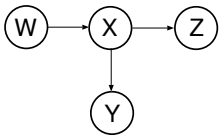
$P(z \mid x)$	z: a	b	c
x: a	1/3	0	2/3
b	1/4	1/2	1/4
c	0	3/5	2/5

a) Representar gráficamente la red

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

c) Obtener una expresión simplificada de $P(W \mid X,Y,Z)$, y calcular $P(W=\text{c} \mid X=\text{b},\ Y=\text{b},\ Z=\text{b})$

a) Representación gráfica de la red:



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

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

$$P(X=\text{b},\ Y=\text{b},\ Z=\text{b} \mid W=\text{b}) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$$

c) Expresión simplificada de $P(W \mid X,Y,Z)$:

$$P(W \mid X,Y,Z) = \frac{P(W,X,Y,Z)}{P(X,Y,Z)} = \frac{P(W)\ P(X \mid W)\ \cancel{P(Y \mid X)\ P(Z \mid X)}}{\sum_{w\in\{\text{a,b,c}\}} P(W=w)P(X \mid W=w)\ \cancel{P(Y \mid X)\ P(Z \mid X)}}$$
$$= \frac{\cancel{(1/3)}\,P(X \mid W)}{\cancel{(1/3)}\sum_{w\in\{\text{a,b,c}\}} P(X \mid W=w)} = \frac{P(X \mid W)}{\sum_{w\in\{\text{a,b,c}\}} P(X \mid W=w)}$$

$$P(W=\text{c} \mid X=\text{b},\ Y=\text{b},\ Z=\text{b}) = \frac{P(X=\text{b} \mid W=\text{c})}{\sum_{w\in\{\text{a,b,c}\}} P(X=\text{b} \mid W=w)} = \frac{3/5}{0+1/2+3/5} = \frac{6}{11}$$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 25 de enero de 2017

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 [C] Sea S un conjunto de 1000 datos supervisados o etiquetados. En el diseño de un sistema de reconocimiento de formas se utiliza el método de *validación cruzada en 10 bloques* y se obtienen los errores siguientes: (2,4,2,3,0,7,4,4,4,1). Indicar cual de las siguientes afirmaciones es incorrecta.

A) El error estimado es $p_e = 3.1\,\%$
B) El intervalo de confianza al 95 % es ± 0.9
C) El test efectivo es de 100 muestras
D) El tamaño de entrenamiento efectivo es de 900 muestras

2 [D] En el problema de aprendizaje de modelos probabilísticos con variables observables $\boldsymbol{x}_n$ y latentes $\boldsymbol{z}_n$

$$L_S(\boldsymbol{\Theta}) \;=\; \sum_{n=1}^{N} \log \left(\sum_{\boldsymbol{z}_n} P(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\Theta}) \right)$$

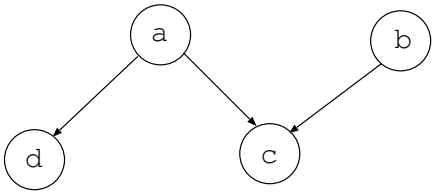
se utiliza la técnica esperanza-maximización (EM). Indicar cuál de las siguientes afirmaciones es cierta:

A) En cada iteración, el paso E consiste en obtener una estimación de todas las variables $\boldsymbol{x}_n$ y $\boldsymbol{z}_n$, y el paso M, obtener los parámetros óptimos de $\boldsymbol{\Theta}$ utilizando la estimación de las variables $\boldsymbol{x}_n$ y $\boldsymbol{z}_n$ obtenidas en el paso E.
B) En cada iteración, el paso E consiste en obtener los valores de las variables latentes $\boldsymbol{z}_n$ que maximizan la función objetivo $L_S(\boldsymbol{\Theta})$, y el paso M, obtener los parámetros óptimos de $\boldsymbol{\Theta}$ utilizando la estimación de las variables $\boldsymbol{z}_n$ obtenidas en el paso E.
C) En cada iteración, el paso E consiste en obtener los valores de todas las variables $\boldsymbol{x}_n$ y $\boldsymbol{z}_n$ que maximizan la función $L_S(\boldsymbol{\Theta})$,, y el paso M, obtener los parámetros óptimos de $\boldsymbol{\Theta}$ utilizando la estimación de las variables $\boldsymbol{x}_n$ y $\boldsymbol{z}_n$ obtenidas en el paso E.
D) En cada iteración, el paso E consiste en obtener una estimación de las variables latentes $\boldsymbol{z}_n$, y el paso M, obtener los parámetros óptimos de $\boldsymbol{\Theta}$ utilizando la estimación de las variables $\boldsymbol{z}_n$ obtenidas en el paso E.

3 [B] Considerar el aprendizaje mediante máquinas de vectores soportes y márgenes blandos con una muestra de aprendizaje $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ no separable linealmente. Si un multiplicador de Lagrange óptimo α_j^*, asociado a la restricción $c_j\ (\boldsymbol{\theta}^t \boldsymbol{x}_d + \theta_0)\ \geq\ 1 - \zeta_j$, $1 \leq j \leq N$, es cero, entonces la muestra $\boldsymbol{x}_j$ está bien clasificada pero ¿cuál de las siguientes afirmaciones es falsa?:

A) $\zeta_j = 0$
B) Se produce un error de margen
C) No hay error de margen
D) La muestra $\boldsymbol{x}_j$ no es un vector soporte

4 [A] En la red bayesiana



¿cuál de las relaciones siguientes es verdadera?

A) $P(a, b) \;=\; P(a)\ P(b)$
B) $P(a, d) \;=\; P(a)\ P(d)$
C) $P(a, b \mid d) \;=\; P(a \mid d)\ P(b \mid d)$
D) $P(a, c \mid b) \;=\; P(a \mid b)\ P(c \mid b)$

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

La siguiente tabla contine una muestra de entrenamiento linealmente separable en $\mathbb{R}^2$ y los correspondientes multiplicadores de Lagrange óptimos obtenidos al entrenar una máquina de vectores soporte con esta muestra:

i	1	2	3	4	5	6	7
x_{i1}	1	1	2	1	4	3	6
x_{i2}	6	3	3	1	5	4	2
Clase	+1	+1	+1	+1	-1	-1	-1
$\alpha_i^\star$	0.25	0	1.0	0	0	1.25	0

a) Obtener la función discriminante lineal correspondiente

b) Calcular el margen óptimo

b) Obtener la ecuación de la frontera de separación entre clases y representarla gráficamente, junto con los datos de entrenamiento.

c) Clasificar la muestra $(2,6)^t$.

a) Función discriminante lineal (FDL):

- Vector de pesos:

$$\begin{aligned}\theta_1^\star &= +1\cdot 0.25\cdot 1 \ +1\cdot 1.0\cdot 2 \ -1\cdot 1.25\cdot 3 = -1.5\\ \theta_2^\star &= +1\cdot 0.25\cdot 6 \ +1\cdot 1.0\cdot 3 \ -1\cdot 1.25\cdot 4 = -0.5\end{aligned}$$

- Peso umbral (con el primer vector de entrenamiento): $\theta_0^\star = (+1) - (-1.5\cdot 1 - 0.5\cdot 6) = 5.5$
- FDL: $\phi(\boldsymbol{x}) = -1.5\ x_1 - 0.5\ x_2 + 5.5$

c) Margen óptimo:

$$\frac{2}{\|\boldsymbol{\theta}^\star\|} = \frac{2}{\sqrt{(-1.5)^2+(-0.5)^2}} = 1.265$$

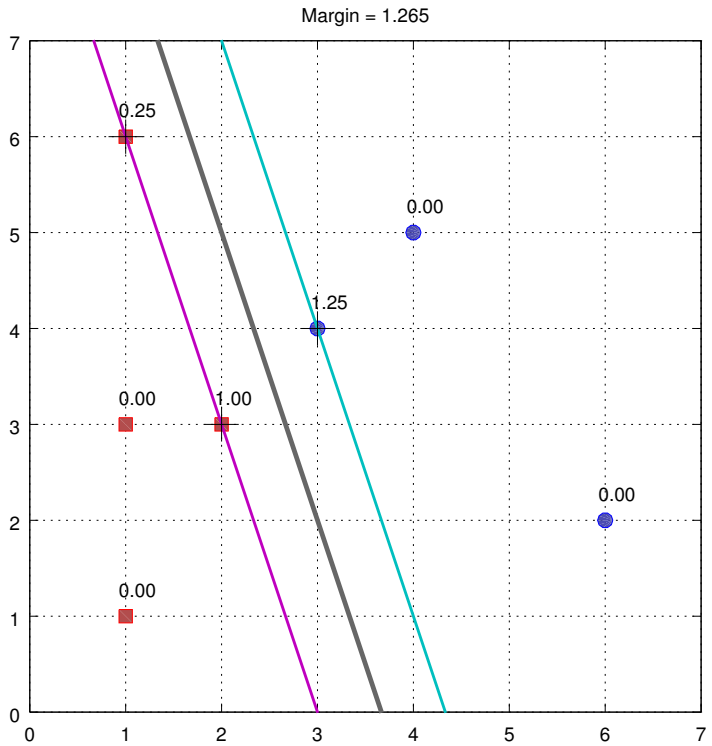
Alternativamente:

$$2\left(\sum_{n\in\mathcal{V}}\alpha_n^\star\right)^{-1/2} = \frac{2}{\sqrt{0.25+1.0+1.25}} = 1.265$$

b) Ecuación de la frontera de decisión:

$$-1.5\ x_1 - 0.5\ x_2 + 5.5 = 0 \quad\Rightarrow\quad x_2 = -3x_1 + 11$$

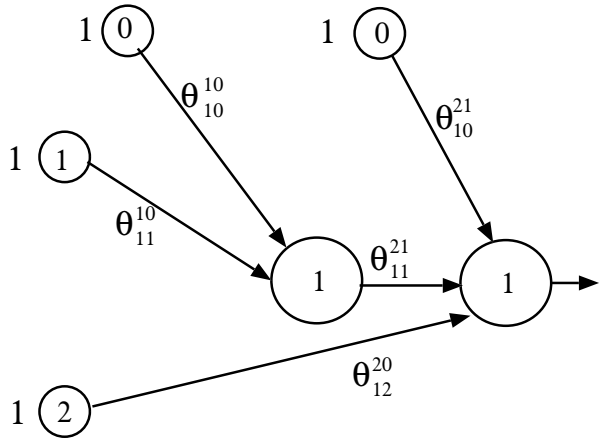
Representación gráfica:



d) Clasificación de la muestra $(2,6)$: $\phi(2,6) = -1.5\cdot 2 - 0.5\cdot 6 + 5.5 = -0.5 < 0 \quad\Rightarrow\quad$ clase $= -1$

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

La red hacia adelante ("feedforward") de la figura se utiliza para resolver un problema de regresión, con función de activación de los nodos de la capa de salida y de la capa oculta de tipo *tangente hiperbólica*, y factor de aprendizaje $\rho = 1.0$.



Dados unos pesos iniciales $\theta_{10}^{10} = \theta_{11}^{10} = \theta_{12}^{20} = \theta_{10}^{21} = \theta_{11}^{21} = 1.0$, un vector de entrada $\boldsymbol{x}^t = (1, 1)$ y su valor deseado de salida $t = +1$, Calcular:

a) las salidas de todas las unidades

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

c) Los nuevos valores de los pesos de las conexiones

a) Las salidas de todas las unidades

$\phi_1^1 = \theta_{10}^{10} + \theta_{11}^{10}\ x_1 = 2.0$ $\qquad$ $s_1^1 = \dfrac{\exp(\phi_1^1) - \exp(-\phi_1^1)}{\exp(\phi_1^1) + \exp(-\phi_1^1)} = 0.96402$

$\phi_1^2 = \theta_{10}^{21} + \theta_{12}^{20}\ x_2 + \theta_{11}^{21}\ s_1^1 = 2.96402$ $\qquad$ $s_1^2 = \dfrac{\exp(\phi_1^2) - \exp(-\phi_1^2)}{\exp(\phi_1^2) + \exp(-\phi_1^2)} = 0.99469$

b) El error en la capa de salida es:

$\delta_1^2 = (t_1 - s_1^2)\ (1 - (s_1^2)^2) = 0.0000562$

El error en la capa de oculta es:

$\delta_1^1 = (\delta_1^2\ \theta_{11}^{21})\ (1 - (s_1^1)^2) = 0.000004$

c) Los nuevos pesos son:

$\theta_{10}^{21} = \theta_{10}^{21} + \rho\ \delta_1^2\ (+1) = 1.0 + 1.0 * 0.0000562 * 1.0 = 1.0000562$

$\theta_{11}^{21} = \theta_{11}^{21} + \rho\ \delta_1^2\ s_1^2 = 1.0 + 1.0 * 0.0000562 * 0.96402 = 1.0000542$

$\theta_{12}^{20} = \theta_{12}^{20} + \rho\ \delta_1^2\ x_2 = 1.0 + 1.0 * 0.0000562 * 1.0 = 1.0000562$

$\theta_{10}^{10} = \theta_{10}^{10} + \rho\ \delta_1^1\ (+1) = 1.0 + 1.0 * 0.000004 * 1.0 = 1.0000040$

$\theta_{11}^{10} = \theta_{11}^{10} + \rho\ \delta_1^1\ x_1 = 1.0 + 1.0 * 0.000004 * 1.0 = 1.0000040$

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

Considerar la red bayesiana $\mathcal{R}$ definida como $P(W,X,Y,Z) = P(X)\ P(Y \mid X)\ P(Z \mid X)\ P(W \mid Y,Z)$, cuyas variables aleatorias, W, X, Y, Z, toman valores en en el conjunto {a,b}. Las distribuciones de probabilidad asociadas son como sigue:

- $P(X)$ es uniforme: $P(X = \text{a}) = P(X = \text{b})$,
- $P(W \mid Y,Z)$, $P(Y \mid X)$ y $P(Z \mid X)$ vienen dadas en las siguientes tablas:

$P(w \mid y,z)$		w: a	b
Y: a	Z: a	1/2	1/2
a	b	1/4	3/4
b	a	1/5	4/5
b	b	3/5	2/5

$P(y \mid x)$	y: a	b
x: a	1/3	2/3
b	1/4	3/4

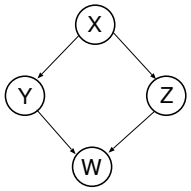
$P(z \mid x)$	z: a	b
x: a	1/3	2/3
b	1/4	3/4

a) Representar gráficamente la red

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

c) Obtener una expresión simplificada de $P(X \mid W,Y,Z)$, y calcular $P(X = \text{b} \mid W = \text{b},\ Y = \text{b},\ Z = \text{b})$

a) Representación gráfica de la red:



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

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

$$P(W = \text{b},\ Y = \text{b},\ Z = \text{b} \mid X = \text{b}) = \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{2}{5} = \frac{9}{40} = 0.225$$

c) Expresión simplificada de $P(X \mid W,Y,Z)$:

$$P(X \mid W,Y,Z) = \frac{P(W,X,Y,Z)}{P(W,Y,Z)} = \frac{P(X)\ P(Y \mid X)\ P(Z \mid X)\ \cancel{P(W \mid Y,Z)}}{\sum_{x\in\{\text{a,b}\}} P(X = x)P(Y \mid X = x)P(Z \mid X = x)\ \cancel{P(W \mid Y,Z)}}$$
$$= \frac{\cancel{(1/2)}\,P(Y \mid X)P(Z \mid X)}{\cancel{(1/2)}\sum_{x\in\{\text{a,b}\}} P(Y \mid X = x)P(Z \mid X = x)} = \frac{P(Y \mid X)P(Z \mid X)}{\sum_{x\in\{\text{a,b}\}} P(Y \mid X = x)P(Z \mid X = x)}$$

$$P(X = \text{b} \mid W = \text{b},\ Y = \text{b},\ Z = \text{b}) = \frac{P(Y = \text{b} \mid X = \text{b})P(Z = \text{b} \mid X = \text{b})}{\sum_{x\in\{\text{a,b}\}} P(Y = \text{b} \mid X = x)P(Z = \text{b} \mid X = x)} = \frac{3/4 \cdot 3/4}{3/4 \cdot 3/4 + 2/3 \cdot 2/3} = 0.5586$$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 10 de enero de 2018

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 [C] Sea S un conjunto de 1000 datos supervisados o etiquetados. Para el diseño de un sistema de reconocimiento de formas, se utilizan datos de S tanto para aprender los parámetros del modelo de reconocimiento, $\mathcal{M}$, como para estimar el error de reconocimiento dicho modelo. Para ello se ha utilizado el método de *validación cruzada en* 10 *bloques*, obteniéndose 4, 0, 4, 3, 4, 0, 3, 5, 4, 3 errores. Indicar cual de las siguientes afirmaciones es *correcta*.

A) El test efectivo es de 100 muestras, el conjunto de entrenamiento efectivo es de 900 muestras, el error empírico $\hat{p}$ es del 3 % y el intervalo de confianza es de 3.0 ± 0.1 %

B) El test efectivo es de 1000 muestras, el conjunto de entrenamiento efectivo es de 1000 muestras, el error empírico $\hat{p}$ es del 3 % y el intervalo de confianza es de 3 ± 1 %

C) El test efectivo es de 1000 muestras, el conjunto de entrenamiento efectivo es de 900 muestras, el error empírico $\hat{p}$ es del 3 % y el intervalo de confianza es de 3 ± 1 %

D) El test efectivo es de 1000 muestras, el conjunto de entrenamiento efectivo es de 900 muestras, el error empírico $\hat{p}$ es del 6 % y el intervalo de confianza es de 6 ± 1 %

2 [A] Sea $S = \{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $c_n \in \{+1, -1\}$ una muestra de entrenamiento. El algoritmo perceptrón muestra a muestra ("online") trata de encontrar una solución $\hat{\boldsymbol{\theta}}$ que satisfaga el sistema de N inecuaciones. Si queremos una solución con *margen*, esto es encontrar una solución $\hat{\boldsymbol{\theta}}$ que satisfaga el sistema de N inecuaciones: $c_n\ \boldsymbol{\theta}^t \boldsymbol{x}_n \geq b, \quad 1 \leq n \leq N$, para $b \in \mathbb{R}^{\geq 0}$. Indicar cuál de los siguientes algoritmos implementa la solución con margen, partiendo de que $\boldsymbol{\theta}(1) =$ arbitrario:

A) $\boldsymbol{\theta}(k+1) = \begin{cases} \boldsymbol{\theta}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) \geq b \\ \boldsymbol{\theta}(k) + \rho_k\ c(k)\ \boldsymbol{x}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) < b \end{cases}$

B) $\boldsymbol{\theta}(k+1) = \begin{cases} \boldsymbol{\theta}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) < b \\ \boldsymbol{\theta}(k) + \rho_k\ c(k)\ \boldsymbol{x}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) \geq b \end{cases}$

C) $\boldsymbol{\theta}(k+1) = \begin{cases} b\ \boldsymbol{\theta}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) \geq 0 \\ b\ \boldsymbol{\theta}(k) + \rho_k\ c(k)\ \boldsymbol{x}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) < 0 \end{cases}$

D) $\boldsymbol{\theta}(k+1) = \begin{cases} \boldsymbol{\theta}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) \geq b \\ \boldsymbol{\theta}(k) + \rho_k\ b\ c(k)\ \boldsymbol{x}(k) & c(k)\ \boldsymbol{\theta}^t \boldsymbol{x}(k) < b \end{cases}$

3 [A] Las siguientes afirmaciones se refieren al método Esperanza Maximización (EM) aplicado a una muestra de entrenamiento S. Identificar cuál de ellas es *correcta*:

A) EM es útil para estimar valores máximo-verosímiles de los parámetros de modelos estadísticos a partir de S cuando hay variables latentes o ocultos.
B) EM no se puede aplicar en la estimación de valores máximo-verosímiles de los parámetros de modelos estadísticos a partir de S cuando no hay variables latentes o ocultas.
C) La rapidez de convergencia de EM puede mejorarse eligiendo un factor de aprendizaje adecuado para S.
D) La rapidez de convergencia de EM siempre puede mejorarse inicializando los parámetros a cero.

4 [D] En la red bayesiana cuya distribución conjunta es $P(x_1, x_2, x_3, x_4) = P(x_1)\ P(x_2 \mid x_1)\ P(x_3 \mid x_1)\ P(x_4 \mid x_2, x_3)$ ¿cuál de las relaciones siguientes es *falsa* en general?

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

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

Para el aprendizaje de una máquina de vectores soporte se dispone de la siguiente muestra de entrenamiento linealmente separable:

$S = \{((6,6),-1),((6,2),-1),((1,6),+1),((2,2),+1),((2,3),+1),((4,2),-1),((3,5),-1),((3,4),-1),((1,2),+1),((2,1),+1)\}$

Los multiplicadores de Lagrange óptimos son: $\boldsymbol{\alpha}^* = (0.0, 0.0, 0.25, 0.0, 1.0, 0.0, 0.0, 1.25, 0.0, 0.0)^t$.

a) Obtener la función discriminante lineal correspondiente

b) Obtener la ecuación de la frontera de decisión entre clases y representar gráficamente los puntos de entrenamiento y dicha frontera.

c) Calcular el margen óptimo

d) Clasificar la muestra $(3,1)$.

a) Función discriminante lineal (FDL):

- Vector de pesos:

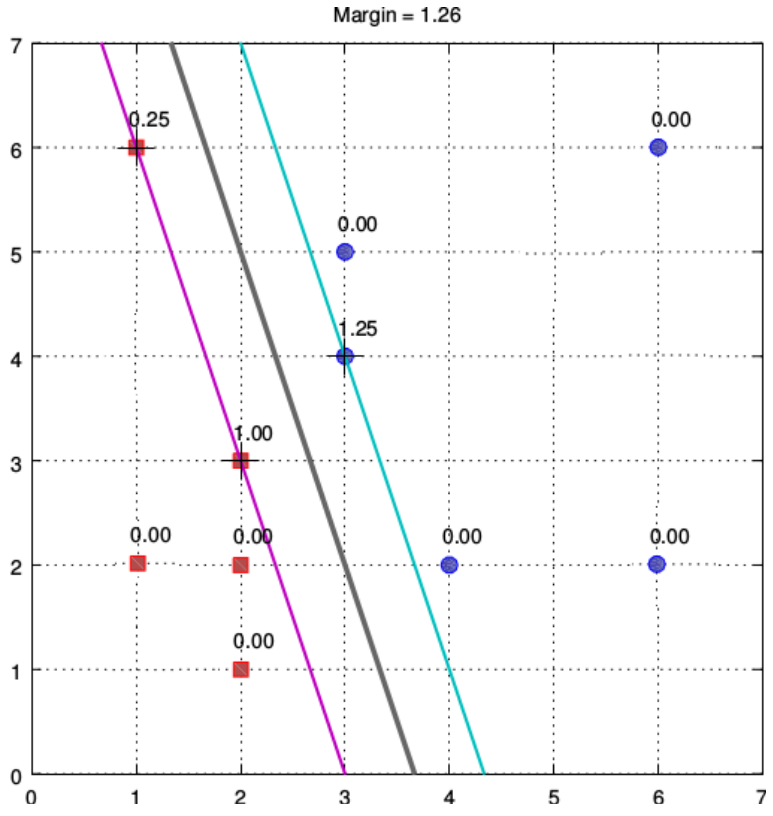
$$\theta_1^\star = +1\cdot 0.25\cdot 1 \; +1\cdot 1.0\cdot 2 \; -1\cdot 1.25\cdot 3 = -1.5$$
$$\theta_2^\star = +1\cdot 0.25\cdot 6 \; +1\cdot 1.0\cdot 3 \; -1\cdot 1.25\cdot 4 = -0.5$$

- Peso umbral (con el tercer vector de entrenamiento): $\theta_0^* = (+1) - (-1.5\cdot 1 - 0.5\cdot 6) = 5.5$
- FDL: $\phi(\boldsymbol{x}) = -1.5\; x_1 - 0.5\; x_2 + 5.5$

b) Ecuación de la frontera de decisión:

$$-1.5\; x_1 - 0.5\; x_2 + 5.5 = 0 \quad \Rightarrow \quad x_2 = -3x_1 + 11$$

Representación gráfica:



c) Margen óptimo:

$$\frac{2}{\|\boldsymbol{\theta}^\star\|} = \frac{2}{\sqrt{(-1.5)^2 + (-0.5)^2}} = 1.265$$

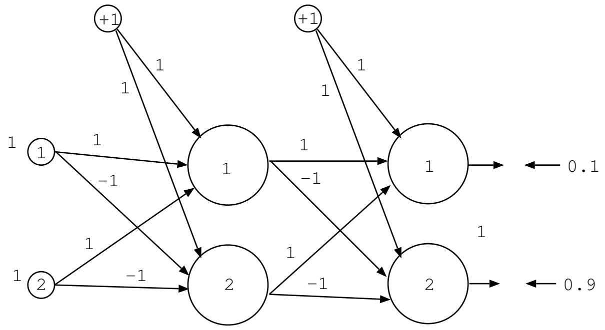
Alternativamente:

$$2\left(\sum_{n\in\mathcal{V}} \alpha_n^\star\right)^{-1/2} = \frac{2}{\sqrt{0.25 + 1.0 + 1.25}} = 1.265$$

d) Clasificación de la muestra $(3,1)$: $\phi(3,1) = -1.5\cdot 3 - 0.5\cdot 1 + 5.5 = 0.5 > 0 \quad \Rightarrow \quad \text{clase} = +1$

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

La red hacia adelante (“feedforward”) de la figura se utiliza para resolver un problema de regresión, con función de activación de los nodos de la capa de salida y de la capa oculta de tipo *sigmoid*, y factor de aprendizaje $\rho = 1.0$.



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

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

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

Capa de salida

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

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

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 $P(W,X,Y,Z) = P(W)\ P(X \mid W)\ P(Y \mid W)\ P(Z \mid X)$, cuyas variables aleatorias, W, X, Y, Z, toman valores en en el conjunto {a,b,c}. Las distribuciones de probabilidad asociadas son como sigue:

- $P(W)$ es uniforme: $P(W=\text{a}) = P(W=\text{b}) = P(W=\text{c})$,
- $P(X \mid W)$, $P(Y \mid W)$ y $P(Z \mid X)$ vienen dadas en las siguientes tablas:

$P(x \mid w)$	x: a	b	c
w: a	1/2	0	1/2
b	1/4	1/2	1/4
c	1/5	3/5	1/5

$P(y \mid w)$	y: a	b	c
w: a	1/3	0	2/3
b	1/4	1/2	1/4
c	0	3/5	2/5

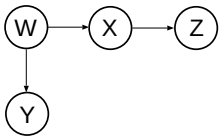
$P(z \mid x)$	z: a	b	c
x: a	1/3	0	2/3
b	1/4	1/2	1/4
c	0	3/5	2/5

a) Representar gráficamente la red

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

c) Obtener una expresión simplificada de $P(W \mid X,Y,Z)$, y calcular $P(W=\text{c} \mid X=\text{b},\ Y=\text{b},\ Z=\text{b})$

a) Representación gráfica de la red:



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

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

$$P(X=\text{b},\ Y=\text{b},\ Z=\text{b} \mid W=\text{c}) = \frac{3}{5}\cdot\frac{3}{5}\cdot\frac{1}{2} = \frac{9}{50} = 0.18$$

c) Expresión simplificada de $P(W \mid X,Y,Z)$:

$$P(W \mid X,Y,Z) = \frac{P(W,X,Y,Z)}{P(X,Y,Z)} = \frac{P(W)\ P(X \mid W)\ P(Y \mid W)\ \cancel{P(Z \mid X)}}{\sum_{w\in\{\text{a,b,c}\}} P(W=w)P(X \mid W=w)\ P(Y \mid W=w)\ \cancel{P(Z \mid X)}}$$
$$= \frac{\cancel{(1/3)}\,P(X \mid W)\ P(Y \mid W)}{\cancel{(1/3)}\sum_{w\in\{\text{a,b,c}\}} P(X \mid W=w)\ P(Y \mid W=w)} = \frac{P(X \mid W)\ P(Y \mid W)}{\sum_{w\in\{\text{a,b,c}\}} P(X \mid W=w)\ P(Y \mid W=w)}$$

$$P(W=\text{c} \mid X=\text{b},\ Y=\text{b},\ Z=\text{b}) = \frac{P(X=\text{b} \mid W=\text{c})\ P(Y=\text{b} \mid W=\text{c})}{\sum_{w\in\{\text{a,b,c}\}} P(X=\text{b} \mid W=w)\ P(Y=\text{b} \mid W=w)}$$
$$= \frac{(3/5)^2}{0+(1/2)^2+(3/5)^2} = \frac{36}{61} \approx 0.59$$

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

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

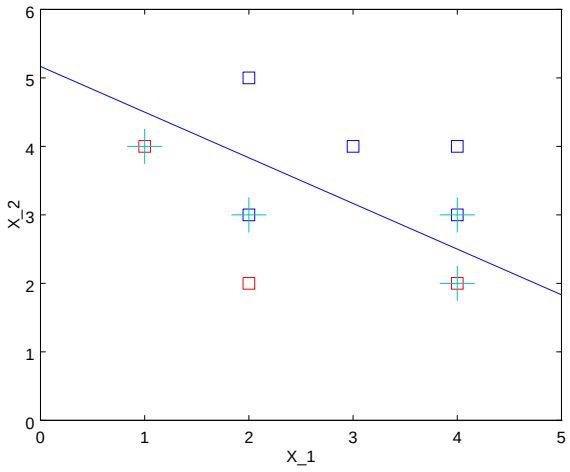
Función discriminante lineal: $\phi(\mathbf{x}) = 10.33 - 1.33\ x_1 - 2.00\ x_2$

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:

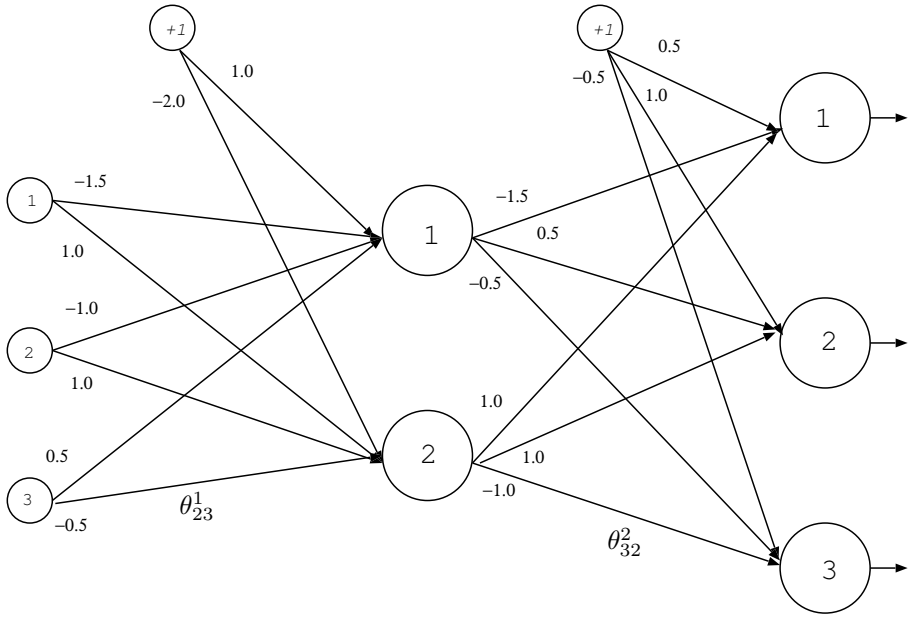


c) Clasificación de la muestra $(2, 4)^t$:

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:

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

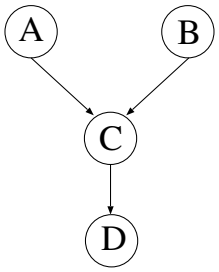
a) Representar gráficamente la red bayesiana

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 $A=0$.

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

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

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

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

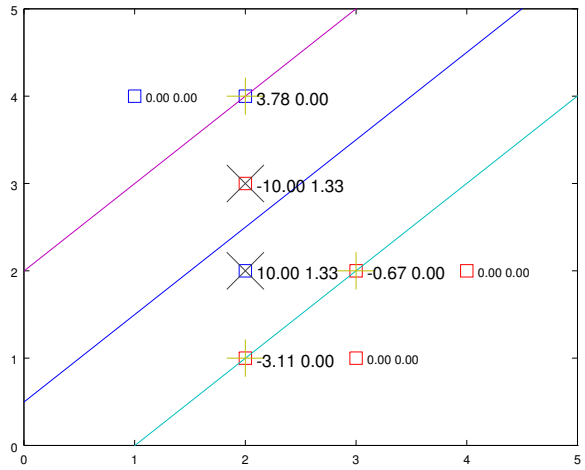
Función discriminante lineal: $\phi(\mathbf{x}) = -0.33 - 0.67\ x_1 + 0.67\ x_2$

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	$t_1 = 0.9$	$t_2 = 0.2$	

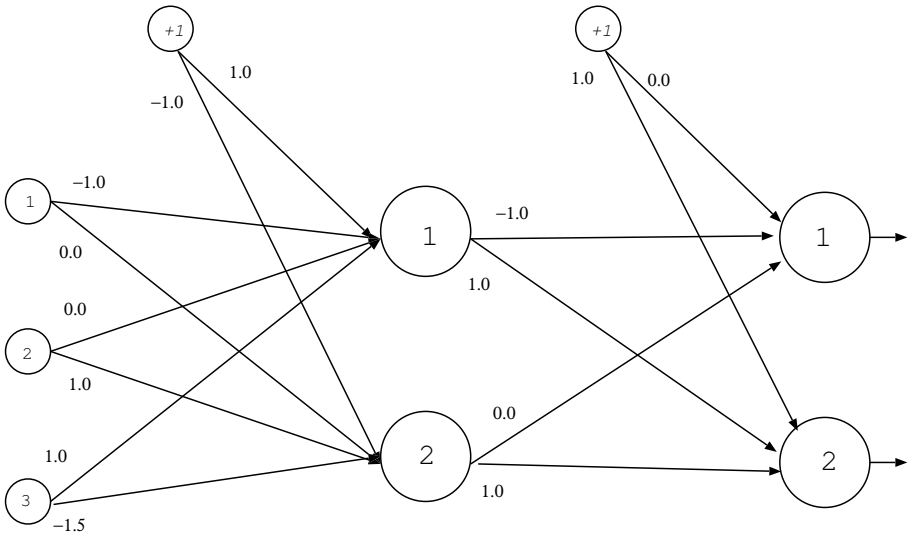
Se pide:

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

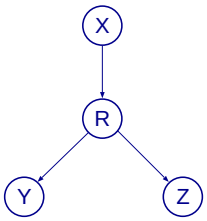
R	"a"	"b"	"c"
1	1/3	0	2/3
2	1/4	1/2	1/4
3	0	3/5	2/5

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

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 23 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] Si aplicamos el método de partición de datos denominado exclusión individual (Leaving One Out) a un conjunto de 1000 muestras, indicar qué opción es *correcta*:

A) La talla de entrenamiento efectiva es de 1000 muestras y la de test de 1000 muestras.
B) La talla de entrenamiento efectiva es de 1000 muestras y la de test de 999 muestras.
C) La talla de entrenamiento efectiva es de 999 muestras y la de test de 999 muestras.
D) La talla de entrenamiento efectiva es de 999 muestras y la de test de 1000 muestras.

2 [A] En un clasificador en 3 clases resulta que la probabilidad a-posteriori de cada clase, y, dada una muestra $x \in \{1,2,3\}$ y la probabilidad a-priori de $x \in \{1,2,3\}$ son:

$P(Y = y \mid X = x)$

	x		
y	1	2	3
A	0.1	0.2	0.5
B	0.6	0.2	0.4
C	0.3	0.6	0.1

x	$P(X = x)$
1	0.2
2	0.3
3	0.5

Indicar cuál es la opción *correcta*:

A) El mínimo riesgo total es 0.45
B) El mínimo riesgo total es 0.55
C) El mínimo riesgo total es 0.5
D) El mínimo riesgo total es 0.4

3 [D] La técnica *descenso por gradiente* aplicada a un problema de optimización da lugar a un algoritmo. Se pide identificar qué afirmación sobre este algoritmo es *falsa*.

A) Se aplica a problemas de minimización
B) Converge asintóticamente en determinadas condiciones
C) Es más rápido si el tamaño del paso de descenso es grande pero no se garantiza la convergencia
D) Siempre converge independientemente del tamaño del paso de descenso

4 [B] Identificar qué afirmación es *cierta* en una máquina de vectores soporte:

A) En el caso de muestras no linealmente separables, los vectores soporte son las muestras que presentan un multiplicador de Lagrange óptimo cero
B) En el caso de muestras no linealmente separable, los vectores soporte son las muestras que presentan un multiplicador de Lagrange óptimo distinto cero aunque estén mal clasificados
C) En el caso de muestras no linealmente separable, los vectores soporte son las muestras que presentan un multiplicador de Lagrange óptimo distinto de cero y tolerancia cero
D) En el caso de muestras no linealmente separable, los vectores soporte son las muestras que presentan un multiplicador de Lagrange óptimo distinto de cero y la tolerancia menor que 1

5 [C] En un modelo gráfico, 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) Solo si el grafo es dirigido, el modelo gráfico define una distribución condicionale en las variables del grafo
B) Si el grafo es no dirigido el número de factores en la distribución conjunta es igual al número de nodos del grafo siempre
C) Cualquier distribución condicional o conjunta en la que participen todas o cualquier subconjunto de las variables de A, se puede deducir a partir de la distribución conjunta definida por G.
D) Si el grafo es dirigido el número de factores en la distribución conjunta es menor que el número de nodos del 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
x_{i1}	1	1	4	4	2	1	3
x_{i2}	4	3	1	3	1	2	2
Clase	+1	+1	−1	−1	−1	+1	+1
$\alpha_i^\star$	0	0	0	5	6	1	10

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) Calcular las tolerancias óptimas.

d) Clasificar la muestra $(3,1)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_4\ \alpha_4^\star\ \mathbf{x_4} + c_5\ \alpha_5^\star\ \mathbf{x_5} + c_6\ \alpha_6^\star\ \mathbf{x_6} + c_7\ \alpha_7^\star\ \mathbf{x_7}$

$\theta_1^* = (-1)\ (5)\ (4) + (-1)\ (6)\ (2) + (+1)\ (1)\ (1) + (+1)\ (10)\ (3) = -1.0$

$\theta_2^* = (-1)\ (5)\ (3) + (-1)\ (6)\ (1) + (+1)\ (1)\ (2) + (+1)\ (10)\ (2) = +1.0$

Usando el vector soporte $\mathbf{x_5}$ (que verifica la condición : $0 < \alpha_5^* < C = 10$)

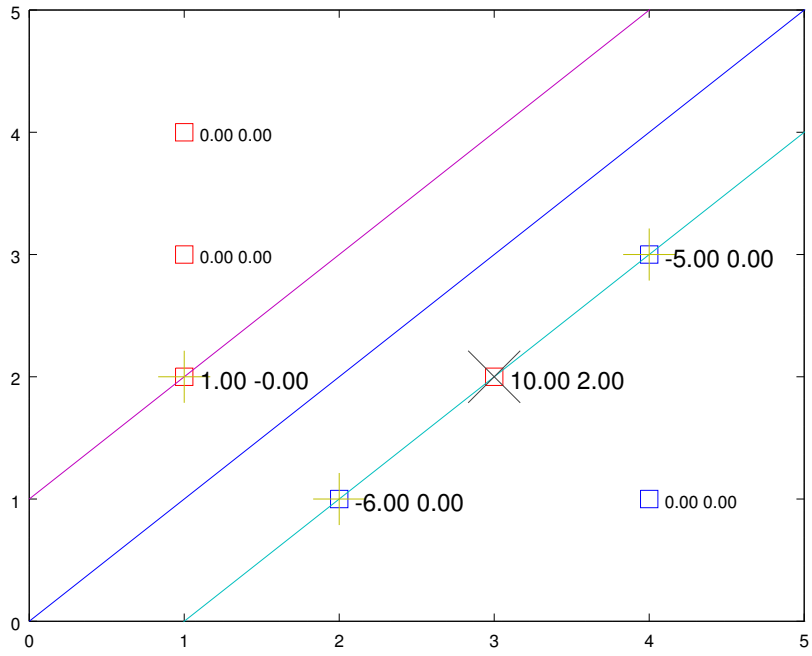
$\theta_0^\star = c_5 - \boldsymbol{\theta}^{\star t}\mathbf{x_5} = -1 - ((-1.0)\ (2) + (1.0)\ (1)) = 0.0$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $(-1.0)\ x_1 + (1.0)\ x_2 + 0.0 = 0.0$

Loss vectores soporte son: $(4,3)^t, (2,1)^t, (1,2)^t, (3,2)^t$.

Representación gráfica:



c) Las tolerancias de todas las muestras excepto la séptima son cero y la séptima que es $\zeta_7^* = 1 - c_7\ (\boldsymbol{\theta}^{*t}\ \boldsymbol{x}_7 + \theta_0^*) = 2.0$;

d) Clasificación de la muestra $(3,1)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 1 + \theta_2^*\ 1 = -2.0 < 0 \Rightarrow$ clase -1.

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

La solución para un determinado problema de regresión viene dado por el 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:

$$\boldsymbol{\theta}_1^1 = (1.0, -1.0, 0.0, 1.0)^t \qquad \boldsymbol{\theta}_1^2 = (0.0, 0.0)^t \qquad \boldsymbol{\theta}_2^2 = (-1.0, -1.0)^t \qquad \boldsymbol{\theta}_3^2 = (1.0, 1.0)^t$$

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	$s_1^1 = 0.5$		
Las salidas de la capa de salida	$s_1^2 = 0.5$	$s_2^2 = 0.182$	$s_3^2 = 0.817$
Los valores deseados de la capa de salida	$t_1 = 0.1$	$t_2 = 0.9$	$t_3 = -0.9$

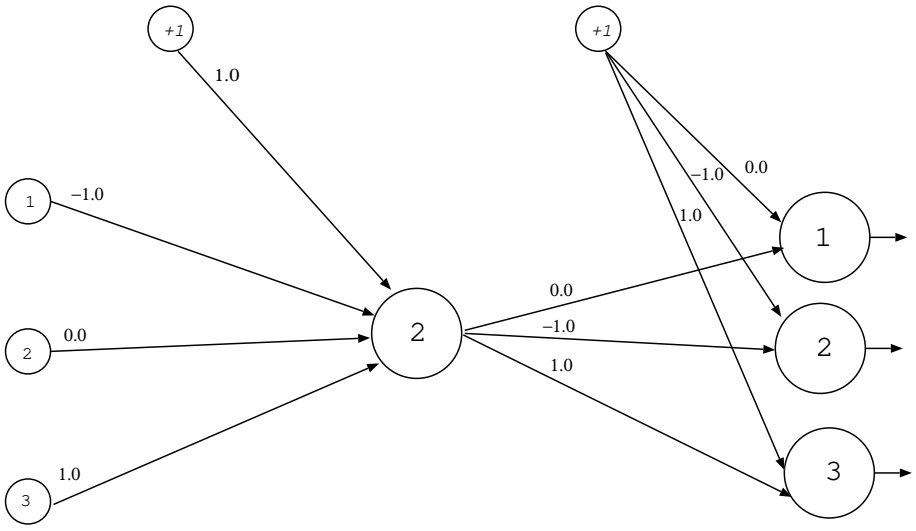
Se pide:

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 $\theta_{3,1}^2$ y $\theta_{1,3}^1$ asumiendo que el factor de aprendizaje ρ es 1.0

a) Dibujo del perceptrón multicapa



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

$\delta_1^2 = (t_1 - s_1^2)\ s_1^2\ (1 - s_1^2) = -0.1$
$\delta_2^2 = (t_2 - s_2^2)\ s_2^2\ (1 - s_2^2) = 0.107$
$\delta_3^2 = (t_3 - s_3^2)\ s_3^2\ (1 - s_3^2) = -0.256$

Errores en la capa de oculta:

$\delta_1^1 = (\delta_1^2\ \theta_{11}^2 + \delta_2^2\ \theta_{21}^2 + \delta_3^2\ \theta_{31}^2) s_1^1\ (1 - s_1^1) = -0.091$

c) Nuevo peso $\theta_{3,1}^2 = \theta_{3,1}^2 + \rho\ \delta_3^2\ s_1^1\ =\ 0.872$

Nuevo peso $\theta_{1,3}^1 = \theta_{1,3}^1 + \rho\ \delta_1^1\ x_3\ =\ 1.091$

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 $\{\texttt{"a"}, \texttt{"b"}, \texttt{"c"}\}$. Las distribuciones de probabilidad asociadas son como sigue:

- $P(X)$ es uniforme: $P(X = \texttt{"a"}) = P(X = \texttt{"b"}) = P(X = \texttt{"c"})$
- $P(R \mid X)$ 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

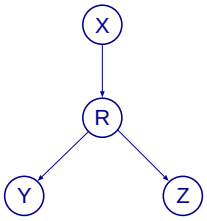
	"a"	"b"	"c"
1	1/3	0	2/3
2	1/4	1/2	1/4
3	0	3/5	2/5

a) Representar gráficamente la red

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

c) Calcular $P(R = 3 \mid X = \texttt{"b"}, Y = \texttt{"b"}, Z = \texttt{"b"})$

a) Representación gráfica de la red:



b) $$P(R \mid X, Y, Z) = \frac{P(R, X, Y, Z)}{P(X, Y, Z)} = \frac{P(X)\ P(R \mid X)\ P(Y \mid R)\ P(Z \mid R)}{\sum_{r=1}^{3} P(X)\ P(R = r \mid X)\ P(Y \mid R = r)\ P(Z \mid R = r)}$$
$$= \frac{P(R \mid X)\ P(Y \mid R)\ P(Z \mid R)}{\sum_{r=1}^{3} P(R = r \mid X)\ P(Y \mid R = r)\ P(Z \mid R = r)}$$
$$= \frac{\frac{1}{3}\ P(Y \mid R)\ P(Z \mid R)}{\sum_{r=1}^{3} \frac{1}{3}\ P(Y \mid R = r)\ P(Z \mid R = r)}$$
$$= \frac{P(Y \mid R)\ P(Z \mid R)}{\sum_{r=1}^{3} P(Y \mid R = r)\ P(Z \mid R = r)}$$

c) $$P(R = 3 \mid X = \texttt{"b"}, Y = \texttt{"b"}, Z = \texttt{"b"}) = \frac{P(Y = \text{"}b\text{"} \mid R = 3)\ P(Z = \text{"}b\text{"} \mid R = 3)}{\sum_{r=1}^{3} P(Y = \text{"}b\text{"} \mid R = r)\ P(Z = \text{"}b\text{"} \mid R = r)}$$
$$= \frac{\frac{3}{5} \cdot \frac{3}{5}}{0 \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{3}{5}} = \frac{9/5}{61/100} = \frac{36}{61} \approx 0.590$$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 7 de enero de 2020

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 [A] 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 que contiene 500 muestras. Se han obtenido un total de 55 errores. Indicar cuál de las afirmaciones siguientes es razonable:

A) Las tallas de entrenamiento y test efectivas son 495 y 500 muestras, respectivamente, y el error estimado es $11.0 \pm 2.7\,\%$
B) Las tallas de entrenamiento y test efectivas son 100 y 400 muestras, respectivamente, y el error estimado es $13.8 \pm 3.0\,\%$
C) Las tallas de entrenamiento y test efectivas son 5 y 495 muestras, respectivamente y el error estimado es $11.1 \pm 2.8\,\%$
D) Ninguna de las anteriores afirmaciones es razonable

2 [A] Considerar la siguiente modificación de la función de Widrow y Hoff

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

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 este caso, ¿cuál de los siguintes gradientes, $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$, es correcto?:

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

3 [A] En el problema de optimización con restricciones

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

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) Para todo i tal que $\alpha_i^\star > 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
B) Para todo i tal que $\alpha_i^\star < 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
C) Si para un i, $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$
D) Para todo i, si $\alpha_i^\star = 0$, entonces $v_i(\boldsymbol{\Theta}^\star) = 0$,

4 [A] 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) El algoritmo *esperanza-maximización* es una alternativa a la técnica de los Multiplicadores de Lagrange en el caso de la estimación de los parámetros de una mezcla de K gaussianas.
B) La verosimilitud del conjunto de entrenamiento, calculada con los parámetros estimados, aumenta en cada iteración del *esperanza-maximización*.
C) En cada iteración, el algoritmo *esperanza-maximización* realiza una estimación de los valores de los pesos de las gaussianas.
D) Los parámetros de la mezcla se estiman adecuadamente mediante un algoritmo de *esperanza-maximización*

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}	4	2	4	2	1	2	3	4
x_{i2}	3	5	2	2	4	3	4	4
Clase	-1	-1	$+1$	$+1$	$+1$	-1	-1	-1
$\alpha_i^\star$	6.22	0	9.11	0	7.11	10	0	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 $(1,1)^t$.

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_5\alpha_5^\star\ \mathbf{x_5} + c_6\alpha_6^\star\ \mathbf{x_6}$

$\theta_1^* = (-1)\ (6.22)\ (4) + (+1)\ (9.11)\ (4) + (+1)\ (7.11)\ (1) + (-1)\ (10.0)\ (2) = -1.33$

$\theta_2^* = (-1)\ (6.22)\ (3) + (+1)\ (9.11)\ (2) + (+1)\ (7.11)\ (4) + (-1)\ (10.0)\ (3) = -2.00$

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

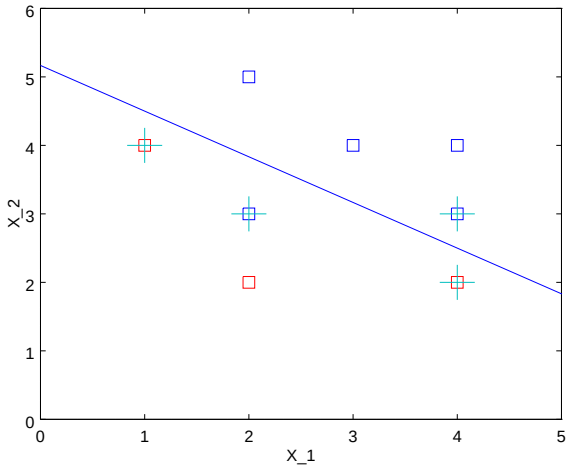
$\theta_0^\star = c_1 - \boldsymbol{\theta}^{\star t}\mathbf{x_1} = -1 - ((-1.33)\ (4) - (2.00)\ (3)) = 10.33$

b) Frontera de separación y representación gráfica:

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

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

Representación gráfica:

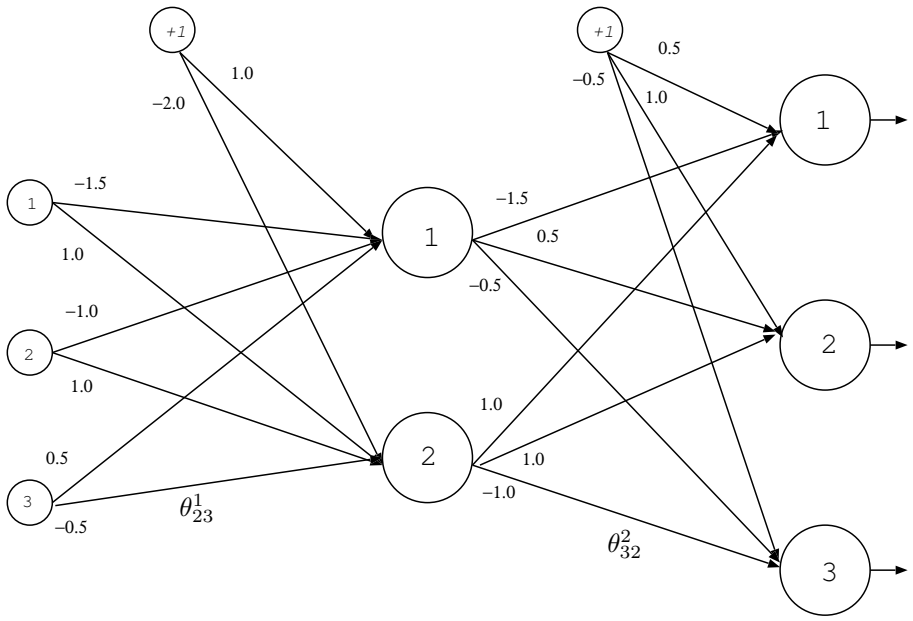


c) Clasificación de la muestra $(1,1)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + \theta_1^*\ 5 + \theta_2^*\ 5 = +7.0 > 0 \Rightarrow$ 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 : $x_1 = 2.0, \quad x_2 = 1.0, \quad x_3 = 2.0$
Los valores deseados de la capa de salida : $t_1 = 2.0, \quad t_2 = 1.0, \quad t_3 = 2.0$
Calcular:

a) Los valores que se obtienen en la unidades ocultas y de salida.

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

c) 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 valores de la capa de salida son:

$s^1_1 = f_s(\theta^1_{1,0} + \theta^1_{1,1}x_1 + \theta^1_{1,2}x_2 + \theta^1_{1,3}x_3) = 0.1192; \qquad s^1_2 = f_s(\theta^1_{2,0} + \theta^1_{2,1}x_1 + \theta^1_{2,2}x_2 + \theta^1_{2,3}x_3) = 0.5$

$s^2_1 = f_s(\theta^2_{1,0}+\theta^2_{1,1}s^1_1+\theta^2_{1,2}s^1_2) = 0.6945; \quad s^2_2 = f_s(\theta^2_{2,0}+\theta^2_{2,1}s^1_1+\theta^2_{2,2}s^1_2) = 0.8263; \quad s^2_3 = f_s(\theta^2_{3,0}+\theta^2_{3,1}s^1_1+\theta^2_{3,2}s^1_2) = 0.2574$

b) Los errores en la capa de salida son:

$\delta^2_1 = (t_1 - s^2_1)\ s^2_1\ (1 - s^2_1) = 0.2770; \qquad \delta^2_2 = (t_2 - s^2_2)\ s^2_2\ (1 - s^2_2) = 0.0249; \qquad \delta^2_3 = (t_3 - s^2_3)\ s^2_3\ (1 - s^2_3) = 0.3331$

Los errores en la capa de oculta son:

$\delta^1_1 = (\delta^2_1\ \theta^2_{11} + \delta^2_2\ \theta^2_{21} + \delta^2_3\ \theta^2_{31})\ s^1_1\ (1 - s^1_1) = -0.0598; \qquad \delta^1_2 = (\delta^2_1\ \theta^2_{12} + \delta^2_2\ \theta^2_{22} + \delta^2_3\ \theta^2_{32})\ s^1_2\ (1 - s^1_2) = -0.0078$

c) El nuevo peso θ^2_{32} es: $\theta^2_{32} = \theta^2_{32} + \rho\ \delta^2_3\ s^1_2 = (-1.0) + (1)\ (0.3331)\ (0.5) = -0.8335$

El nuevo peso θ^1_{23} es: $\theta^1_{23} = \theta^1_{23} + \rho\ \delta^1_2\ x_3 = (-0.5) + (1)\ (-0.0078)\ (2.0) = -0.5156$

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

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

$$
\begin{array}{rlrl}
P(A=1) &= 0.3 & P(A=0) &= 0.7 \\
P(B=1 \mid A=1) &= 0.4 & P(B=0 \mid A=1) &= 0.6 \\
P(B=1 \mid A=0) &= 0.6 & P(B=0 \mid A=0) &= 0.4 \\
P(C=1 \mid B=1) &= 0.2 & P(C=0 \mid B=1) &= 0.8 \\
P(C=1 \mid B=0) &= 0.7 & P(C=0 \mid B=0) &= 0.3 \\
P(D=1 \mid B=1) &= 0.1 & P(D=0 \mid B=1) &= 0.9 \\
P(D=1 \mid B=0) &= 0.5 & P(D=0 \mid B=0) &= 0.5
\end{array}
$$

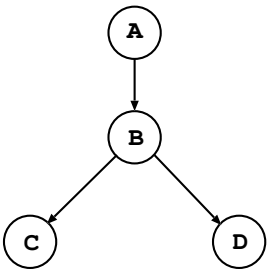
a) Representar gráficamente la red

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

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

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

a) Representación gráfica de la red:



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

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

$$P(B=1,C=1,D=1 \mid A=0) = 0.6 \cdot 0.2 \cdot 0.1 = 0.012$$

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

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

$$P(B=0 \mid A=1,C=1,D=1) = \frac{0.6 \cdot 0.7 \cdot 0.5}{0.6 \cdot 0.7 \cdot 0.5 + 0.4 \cdot 0.2 \cdot 0.1} = 0.9633$$

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

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

$P(B=1 \mid A=1,C=1,D=1) = 1 - 0.9633 = 0.0367$, por tanto la mejor predicción es $B=0$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 21 de enero de 2020

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 [D] Para un problema de clasificación en dos clases, sea $\boldsymbol{\theta}(\mathbf{x};\boldsymbol{\theta}) \stackrel{\text{def}}{=} \boldsymbol{\theta}^t\mathbf{x} + \theta_0$ una función discriminante lineal (FDL) y sea H el hiperplano de decisión definido por $\phi(\mathbf{x};\boldsymbol{\theta}) = 0$. Entre las siguientes supuestas propiedades hay una que es falsa:

A) El valor de $\phi(\mathbf{x};\boldsymbol{\theta})$ es proporcional a la distancia de $\mathbf{x}$ a H
B) La distancia del origen de coordenadas a H es $\dfrac{\theta_0}{\|\boldsymbol{\theta}\|}$
C) H también está definido por un número infinito de FDL $\phi' \neq \phi$
D) Solo hay una única FDL que define a H

2 [A] Se ha evaluado un sistema de Aprendizaje Automático mediante un proceso de *exclusion individual* ("Leaving One Out") usando 1000 muestras etiquetadas. En este proceso se han producido 15 errores en total. Indicar cuál de las afirmaciones siguientes es razonable:

A) La talla de entrenamiento efectiva es de 999 muestras y el error estimado es $1.5\,\% \pm 0.75\,\%$
B) La talla de entrenamiento efectiva es de 1000 muestras y el error estimado es $1.5\,\% \pm 0.15\,\%$
C) La talla de test efectiva es de 999 muestras y el error estimado es $1.5\,\% \pm 0.15\,\%$
D) Las tallas de entrenamiento y de test efectivas son de 1000 muestras y el error estimado es $1.5\,\% \pm 0.75\,\%$

3 [D] Se desea ajustar por mínimos cuadrados la función $f : \mathbb{R}^2 \to R$, definida como: $y = f(\boldsymbol{x}) \stackrel{\text{def}}{=} ax_1x_2 + bx_1 + cx_2$ a una secuencia de N pares entrada-salida: $S = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots, (\boldsymbol{x}_N, y_N))$. La técnica empleada es minimizar por descenso por gradiente la función de error cuadrático:

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

Identifica la afirmación acertada de entre las siguientes:

A) El gradiente es $ax_1 + bx_2 + cx_1x_2$
B) El vector gradiente es: $2\sum_{n=1}^{N}(f(\boldsymbol{x}_n) - y_n) \cdot \boldsymbol{x}_n^t$
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 vector gradiente es: $2\sum_{n=1}^{N}(f(\boldsymbol{x}_n) - y_n) \cdot (x_{n1}x_{n2},\ x_{n1},\ x_{n2}\ )^t$

4 [C] Considerar el aprendizaje mediante máquinas de vectores soportes y márgenes blandos con una muestra de aprendizaje $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ 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, entonces:

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

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
x_{i1}	2	4	3	2	1	4
x_{i2}	3	2	2	2	2	1
Clase	+1	−1	+1	−1	+1	−1
$\alpha_i^\star$	0.67	3.78	10.00	10.00	3.11	0.00

a) Obtener la función discriminante lineal correspondiente y el valor del margen.

b) Representar gráficamente la frontera lineal de separación entre clases, los márgenes y las muestras de entrenamiento, indicando cuáles son vectores soporte.

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

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_1\ \alpha_1^\star\ \mathbf{x_1} + c_2\ \alpha_2^\star\ \mathbf{x_2} + c_3\ \alpha_3^\star\ \mathbf{x_3} + c_4\ \alpha_4^\star\ \mathbf{x_4} + c_5\ \alpha_5^\star\ \mathbf{x_5}$

$\theta_1^* = (+1)\ (0.67)\ (2) + (-1)\ (3.78)\ (4) + (+1)\ (10)\ (3) + (-1)\ (10)\ (2) + (+1)\ (3.11)\ (1) \approx -0.67$

$\theta_2^* = (+1)\ (0.67)\ (3) + (-1)\ (3.78)\ (2) + (+1)\ (10)\ (2) + (-1)\ (10)\ (2) + (+1)\ (3.11)\ (2) \approx +0.67$

Usando el vector soporte $\mathbf{x_2}$ (que verifica la condición : $0 < \alpha_2^* < C = 10$)

$\theta_0^\star = c_2 - \boldsymbol{\theta}^{\star t}\mathbf{x_2} = -1 - ((-0.67)\ (4) + (0.67)\ (2)) = 0.34$

Margen: $\dfrac{2}{\|\boldsymbol{\theta}\|} \approx 2.11$

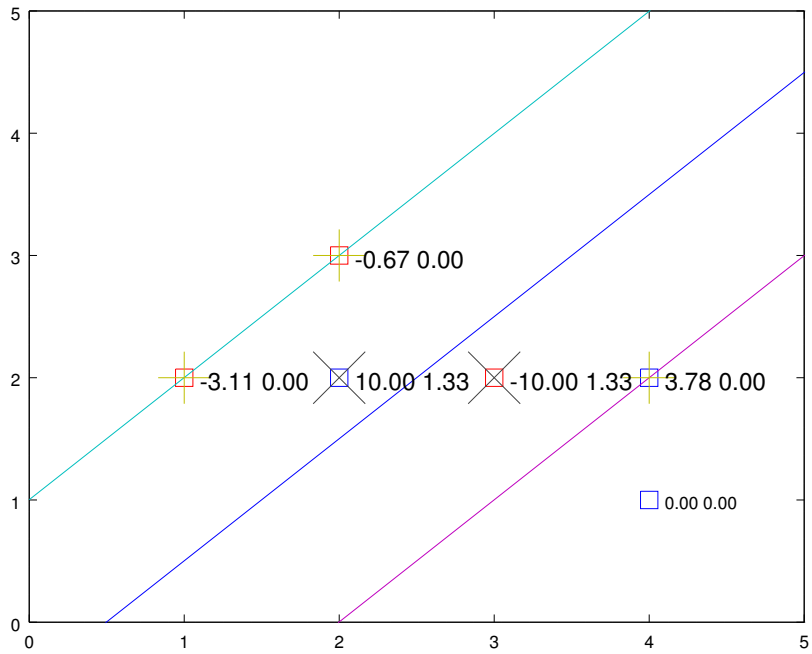
b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $-0.67\ x_1 + 0.67\ x_2 + 0.34 = 0$

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

El margen lo definen las dos rectas paralelas a la frontera de separación, cada una de ellas situada a una distancia de $2.11/2 \approx 1.05$ y cuyas ecuaciones son: $-0.67\ x_1 + 0.67\ x_2 + 0.34 = +1$ y $-0.67\ x_1 + 0.67\ x_2 + 0.34 = -1$

Representación gráfica:



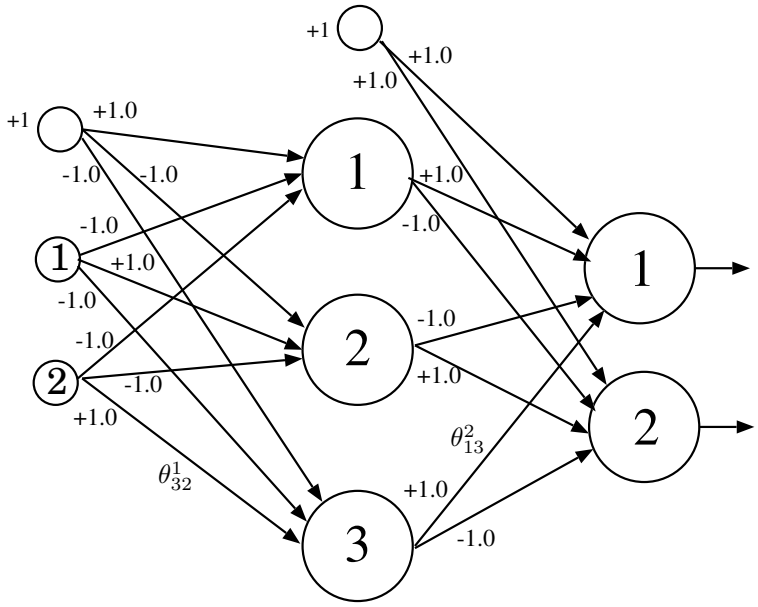
c) Clasificación de la muestra $(4,4)^t$:

El valor de la función discriminante para este vector es:

$\theta_0^* + \theta_1^*\ x_1 + \theta_2^*\ x_2 = +0.34 + (-0.67) * 4 + (0.67) * 4 = +0.34 > 0 \Rightarrow$ 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 función de activación de los nodos de la capa de salida de tipo lineal y de la capa oculta de tipo *sigmoide*, y factor de aprendizaje $\rho = 1.0$.



Dado un vector de entrada $\boldsymbol{x}^t = (+2, +2)$, las salidas de las unidades de la capa de salida son $s^2_1 = 1.0474$ y $s^2_2 = 0.9526$ y las de las unidades ocultas son $s^1_1 = 0.0474$, $s^1_2 = 0.2689$ y $s^1_3 = 0.2689$. Si el valor deseado de salida es $\boldsymbol{t}^t = (+1, 0)$, calcular:

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

b) Los nuevos valores de los pesos de las conexiones θ^1_{32} y θ^2_{13}.

a) Los errores en la capa de salida (función de activación lineal) son:

$\delta^2_1 = (t_1 - s^2_1) = -0.0474$

$\delta^2_2 = (t_2 - s^2_2) = -0.9526$

Los errores en la capa de oculta son:

$\delta^1_1 = (\delta^2_1\, \theta^2_{11} + \delta^2_2\, \theta^2_{21})\, s^1_1\, (1 - s^1_1) = +0.0409;$

$\delta^1_2 = (\delta^2_1\, \theta^2_{12} + \delta^2_2\, \theta^2_{22})\, s^1_2\, (1 - s^1_2) = -0.1780;$

$\delta^1_3 = (\delta^2_1\, \theta^2_{13} + \delta^2_2\, \theta^2_{23})\, s^1_3\, (1 - s^1_3) = +0.1780$

b) El nuevo peso θ^2_{13} es: $\theta^2_{13} = \theta^2_{13} + \rho\, \delta^2_1\, s^1_3 = 0.9872;$

El nuevo peso θ^1_{32} es: $\theta^1_{32} = \theta^1_{32} + \rho\, \delta^1_3\, x_2 = 1.3559$

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

Las variables aleatorias A, B, C, D toman valores en el conjunto $\{0, 1\}$. La distribución de probabilidad conjunta de estas variables viene dada por $P(A, B, C, D) = P(A)\ P(B)\ P(C \mid A, B)\ P(D \mid A, B)$, donde las distribuciones de probabilidad asociadas son:

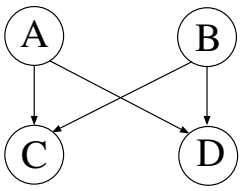
$$
\begin{array}{rr}
P(A=1) = 0.3 & P(A=0) = 0.7 \\
P(B=1) = 0.4 & P(B=0) = 0.6 \\
P(C=1 \mid A=0, B=0) = 0.1 & P(C=0 \mid A=0, B=0) = 0.9 \\
P(C=1 \mid A=0, B=1) = 0.2 & P(C=0 \mid A=0, B=1) = 0.8 \\
P(C=1 \mid A=1, B=0) = 0.3 & P(C=0 \mid A=1, B=0) = 0.7 \\
P(C=1 \mid A=1, B=1) = 0.4 & P(C=0 \mid A=1,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

a) Representar gráficamente la red bayesiana correspondiente

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 $D = 1$.

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

a) Representar gráficamente la red bayesiana correspondiente



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 $D = 1$.

$$
\begin{aligned}
P(A \mid B, C, D) &= \frac{P(A, B, C, D)}{P(B, C, D)} = \frac{P(A)\ \cancel{P(B)}\ P(C \mid A, B)\ P(D \mid A, B)}{\cancel{P(B)}\ \sum_a P(A=a)\ P(C \mid A=a, B)\ P(D \mid A=a, B)} \\
&= \frac{P(A)\ P(C \mid A, B)\ P(D \mid A, B)}{P(A=0)\ P(C \mid A=0, B)\ P(D \mid A=0, B) + P(A=1)\ P(C \mid A=1, B)\ P(D \mid A=1, B)}
\end{aligned}
$$

$$
P(A=1 \mid B=1, C=1, D=1) = \frac{0.3 \cdot 0.4 \cdot 0.6}{0.7 \cdot 0.2 \cdot 0.8 \ +\ 0.3 \cdot 0.4 \cdot 0.6} = 0.391
$$

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

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

$P(A = 0 \mid B = 1, C = 1, D = 1) = 1 - 0.391 = 0.609 \geq 0.391 \Rightarrow$ valor óptimo de A es $a^\star = 0$

Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 18 de enero de 2021

Apellidos: _____ **Nombre:** _____ **Grupo:** ____

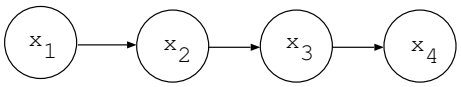
Cuestiones (2 puntos; tiempo estimado: 30 minutos)

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 [C] Si la función de activación en la capa de salida de un perceptrón de dos capas es lineal, las fórmulas que permiten modificar los pesos de dicha capa de salida en el algoritmo BackProp verifican que (solo una respuesta es correcta):

A) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ g(\phi^2_i)\ (1 - g(\phi^2_i))\ s^1_j$
B) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ g(\phi^2_i)\ s^1_j$
C) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ s^1_j$
D) $\Delta\theta^2_{ij} = \rho\ (t_i - s^2_i)\ (1 - g(\phi^2_i))\ s^1_j$

2 [A] En la red bayesiana lineal



¿cuál de las relaciones siguientes es correcta?

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

3 [D] Se ha evaluado un sistema de Aprendizaje Automático mediante la técnica de *validación cruzada en B bloques* ("B-fold Cross Validation") con $B = 10$ y utilizando un conjunto de datos etiquetados que contiene 1000 muestras. Se han obtenido un total de 20 errores. Indicar cuál de las afirmaciones siguientes es correcta:

A) La talla de entrenamiento efectiva es de 1000 muestras y la talla de test efectiva es 1000 muestras.
B) La talla de entrenamiento efectiva es de 900 muestras y el error es del $20.0 \pm 0.2\,\%$
C) La talla de entrenamiento efectiva es de 1000 muestras y y el error es del $20.0 \pm 0.2\,\%$
D) La talla de entrenamiento efectiva es de 900 muestras y el error es del $2.0 \pm 0.9\,\%$

4 [A] 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) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \frac{\lambda}{2}$
B) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + 1$
C) $\displaystyle\sum_{n=1}^{N} \boldsymbol{x}_n + \lambda\ \boldsymbol{\theta}(k)$
D) $\displaystyle\sum_{n=1}^{N} \boldsymbol{\theta}(k)^t \boldsymbol{x}_n + 1$

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	3	4	4	3	1	3	2
x_{i2}	4	2	1	3	1	2	3	3
Clase	+1	+1	−1	−1	−1	+1	+1	−1
$\alpha_i^\star$	0.0	10.0	0.0	6.0	6.4	2.4	10.0	10.0

a) Obtener la función discriminante lineal correspondiente y el valor del margen.

b) Calcular las tolerancias de cada muestra de aprendizaje.

c) Representar gráficamente la frontera lineal de separación entre clases, los márgenes y las muestras de entrenamiento, indicando cuáles son vectores soporte.

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

a) Pesos de la función discriminante:

$\boldsymbol{\theta}^\star = c_2\ \alpha_2^\star\ \mathbf{x_2} + c_4\ \alpha_4^\star\ \mathbf{x_4} + c_5\ \alpha_5^\star\ \mathbf{x_5} + c_5\ \alpha_6^\star\ \mathbf{x_6} + c_5\ \alpha_7^\star\ \mathbf{x_7} + c_5\ \alpha_8^\star\ \mathbf{x_8}$

$\theta_1^* = (+1)\ 10.0\ 3 + (-1)\ 6.0\ 4 + (-1)\ 6.4\ 3 + (+1)\ 2.4\ 1 + (+1)\ 10.0\ 3 + (-1)\ 10\ 2 = -0.8$

$\theta_2^* = (+1)\ 10.0\ 2 + (-1)\ 6.0\ 3 + (-1)\ 6.4\ 1 + (+1)\ 2.4\ 2 + (+1)\ 10.0\ 3 + (-1)\ 10\ 3 = +0.4$

Usando el vector soporte $\mathbf{x_4}$ (que verifica la condición : $0 < \alpha_4^* < C = 10$)

$\theta_0^\star = c_4 - \boldsymbol{\theta}^{\star t}\mathbf{x_4} = -1 - ((-0.8)\ (4) + (0.4)\ (3)) = 1.0$

Margen: $\dfrac{2}{\|\boldsymbol{\theta}\|} \approx 2.23$

b) Calcular las tolerancias de cada muestra de aprendizaje:

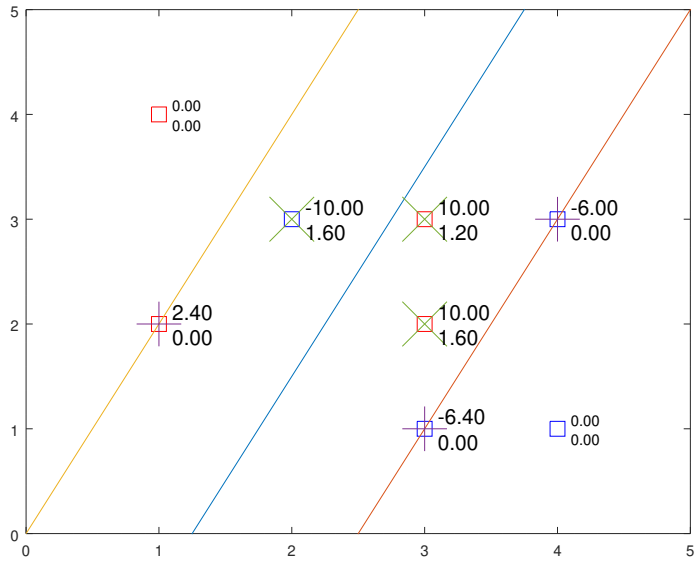
$\zeta_1 = 0.0$; $\zeta_2 = 1 - c_2\ (\boldsymbol{\theta}^{\star t}\mathbf{x_2} + \theta_0^\star) = 1.6$; $\zeta_3 = 0.0$; $\zeta_4 = 0.0$;

$\zeta_5 = 0.0$; $\zeta_6 = 0.0$; $\zeta_7 = 1 - c_7\ (\boldsymbol{\theta}^{\star t}\mathbf{x_7} + \theta_0^\star) = 1.2$; $\zeta_8 = 1 - c_8\ (\boldsymbol{\theta}^{\star t}\mathbf{x_8} + \theta_0^\star) = 1.6$

c) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $-0.8\ x_1 + 0.4\ x_2 + 1.0 = 0$

Los vectores de entrenamiento son todos los de la tabla. De ellos, los vectores soporte son: $(2,3)^t, (4,2)^t, (3,2)^t, (2,2)^t\ (1,2)^t$.
El margen lo definen las dos rectas paralelas a la frontera de separación, cada una de ellas situada a una distancia de $2.23/2 \approx 1.12$ y cuyas ecuaciones son: $-0.8\ x_1 + 0.4\ x_2 + 1.0 = +1$ y $-0.8\ x_1 + 0.4\ x_2 + 1.0 = -1$
Representación gráfica:



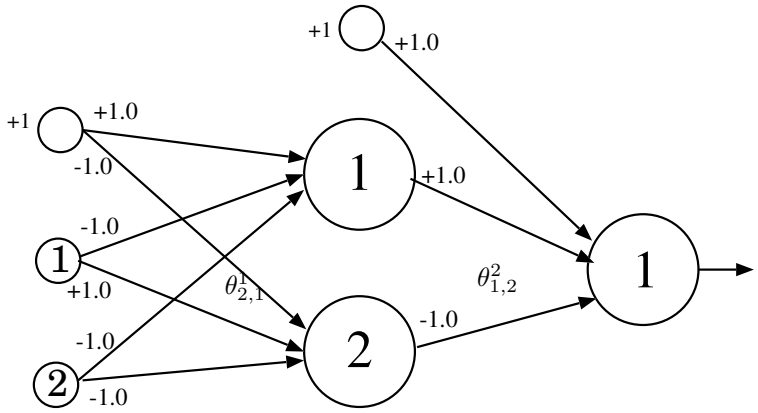
c) Clasificación de la muestra $(4,4)^t$:

El valor de la función discriminante para este vector es:

$\theta_1^*\ x_1 + \theta_2^*\ x_2 + \theta_0^* = (-0.8) * 4 + (0.467) * 4 + 1.0 = -0.6 < 0 \Rightarrow$ 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 función de activación del nodo de la capa de salida de tipo lineal y de los nodos de la capa oculta de tipo *sigmoide*, y factor de aprendizaje $\rho = 1.0$.



Dado un par de entrenamiento $(\boldsymbol{x}^t, t) = ((+2, +2), -1)$, 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 $\theta^2_{1,2}$ y $\theta^1_{2,1}$.

a) Las salidas de la capa oculta son:

$\phi^1_1 = \theta^1_{1,0} + \theta^1_{1,1}\, x_1 + \theta^1_{1,2}\, x_2 = -3 \qquad s^1_1 = f_s(\phi^1_1) = +0.047426$

$\phi^1_2 = \theta^1_{2,0} + \theta^1_{2,1}\, x_1 + \theta^1_{2,2}\, x_2 = -1 \qquad s^1_2 = f_s(\phi^1_1) = +0.268941$

La salida de la capa de salida es:

$\phi^2_1 = \theta^2_{1,0} + \theta^2_{1,1}\, s^1_1 + \theta^2_{1,2}\, s^1_2 = +0.77848 \qquad s^2_1 = f_l(\phi^2_1) = +0.77848$

b) El error en la capa de salida es:

$\delta^2_1 = (t_1 - s^2_1) = -1.7785$

Los errores en la capa oculta son:

$\delta^1_1 = (\delta^2_1\, \theta^2_{1,1})\, f'_S(\phi^1_1) = (\delta^2_1\, \theta^2_{1,1})\, s^1_1\, (1 - s^1_1) = -0.08035$

$\delta^1_2 = (\delta^2_1\, \theta^2_{1,2})\, f'_S(\phi^1_2) = (\delta^2_1\, \theta^2_{1,2})\, s^1_2\, (1 - s^1_2) = +0.34967$

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

$\theta^2_{1,2} = \theta^2_{1,2} + \Delta\theta^2_{1,2} = \theta^2_{1,2} + \rho\, \delta^2_1\, s^1_2 = -1.0 + 1.0\, (-1.7785)\, 0.268941 = -1.47831$

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

$\theta^1_{2,1} = \theta^1_{2,1} + \Delta\theta^1_{2,1} = \theta^1_{2,1} + \rho\, \delta^1_2\, x_1 = +1.0 + 1.0\, (+0.34967)\, 2.0 = +1.69934$

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

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

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

con las correspondientes distribuciones de probabilidad:

$$\begin{array}{cc} P(A=1) \;=\; 0.3 & P(A=0) \;=\; 0.7 \\ P(B=1 \mid A=1) \;=\; 0.4 & P(B=0 \mid A=1) \;=\; 0.6 \\ P(B=1 \mid A=0) \;=\; 0.6 & P(B=0 \mid A=0) \;=\; 0.4 \\ P(C=1 \mid B=1) \;=\; 0.5 & P(C=0 \mid B=1) \;=\; 0.5 \\ P(C=1 \mid B=0) \;=\; 0.3 & P(C=0 \mid B=0) \;=\; 0.7 \\ P(D=1 \mid A=1) \;=\; 0.9 & P(D=0 \mid A=1) \;=\; 0.1 \\ P(D=1 \mid A=0) \;=\; 0.1 & P(D=0 \mid A=0) \;=\; 0.9 \\ P(E=1 \mid D=1) \;=\; 0.2 & P(E=0 \mid D=1) \;=\; 0.8 \\ P(E=1 \mid D=0) \;=\; 0.1 & P(E=0 \mid D=0) \;=\; 0.9 \end{array}$$

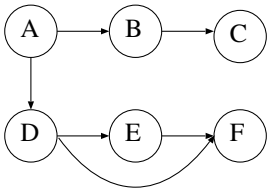
$$\begin{array}{ccc} P(F=x \mid D=0, E=0) \;=\; 0.1 & P(F=y \mid D=0, E=0) \;=\; 0.2 & P(F=z \mid D=0, E=0) \;=\; 0.7 \\ P(F=x \mid D=0, E=1) \;=\; 0.0 & P(F=y \mid D=0, E=1) \;=\; 0.2 & P(F=z \mid D=0, E=1) \;=\; 0.8 \\ P(F=x \mid D=1, E=0) \;=\; 0.3 & P(F=y \mid D=1, E=0) \;=\; 0.3 & P(F=z \mid D=1, E=0) \;=\; 0.4 \\ P(F=x \mid D=1, E=1) \;=\; 0.4 & P(F=y \mid D=1, E=1) \;=\; 0.3 & P(F=z \mid D=1, E=1) \;=\; 0.3 \end{array}$$

a) Representar gráficamente la red bayesiana correspondiente

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

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

a) Representar gráficamente la red bayesiana correspondiente



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

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

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

$$f^\star \;=\; \operatorname*{arg\,max}_{f \in \{x,y,z\}} P(F=f \mid A=1, B=1, C=1, D=1) \;=\; \operatorname*{arg\,max}_{f \in \{x,y,z\}} \sum_e P(E=e \mid D=1)\; P(F=f \mid D=1, E=e)$$

$$P(F=x \mid A=1, B=1, C=1, D=1) \;=\; \sum_e P(E=e \mid D=1)\, P(F=x \mid D=1, E=e) \;=\; 0.2{\cdot}0.4 + 0.8{\cdot}0.3 \;=\; 0.32$$

$$P(F=y \mid A=1, B=1, C=1, D=1) \;=\; \sum_e P(E=e \mid D=1)\, P(F=y \mid D=1, E=e) \;=\; 0.2{\cdot}0.3 + 0.8{\cdot}0.3 \;=\; 0.30$$

$$P(F=z \mid A=1, B=1, C=1, D=1) \;=\; 1 \;-\; 0.32 \;-\; 0.30 \;=\; 0.38$$

El valor óptimo de F es $f^\star = z$

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 [A] En el problema de optimización con restricciones

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

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 [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: $\sum_{k=1}^K \pi_k = 1$.

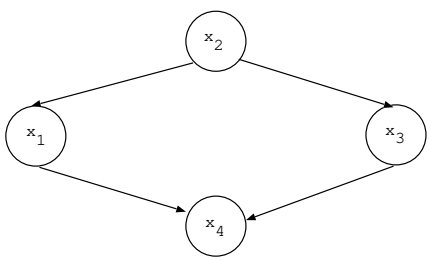
3 [D] Se desea ajustar por mínimos cuadrados la función $f : \mathbb{R} \to \mathbb{R}$, definida como: $y = f(x) \stackrel{\text{def}}{=} ax^2 + bx + c$ a una secuencia de N pares entrada-salida: $S = ((x_1, y_1), (x_2, y_2) \ldots, (x_N, y_N))$. La técnica empleada es minimizar por descenso por gradiente 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 [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 : $0 < \alpha_4^* < C$) $\quad \theta_0^\star = c_4 - \boldsymbol{\theta}^{\star t}\mathbf{x_4} \approx 2.33$

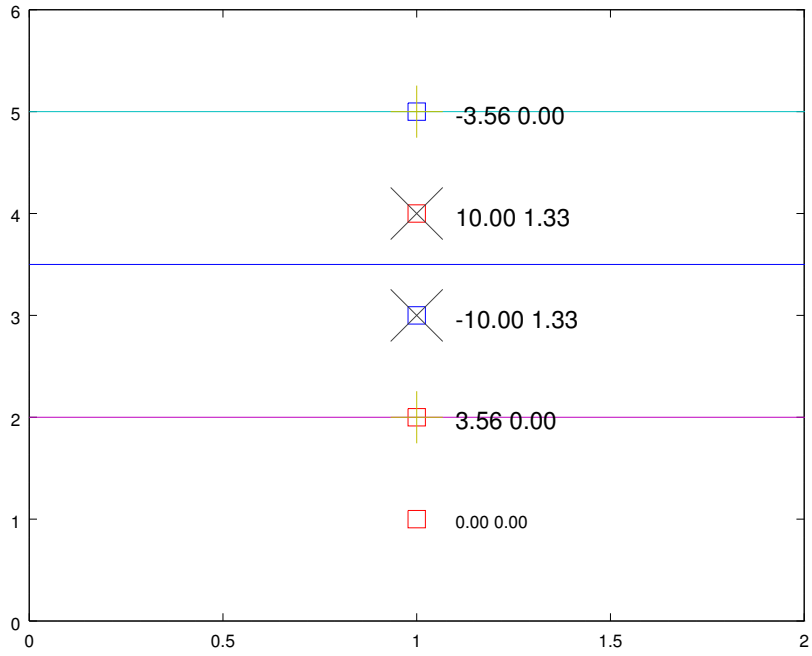
b) Frontera de separación y representación gráfica:

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:

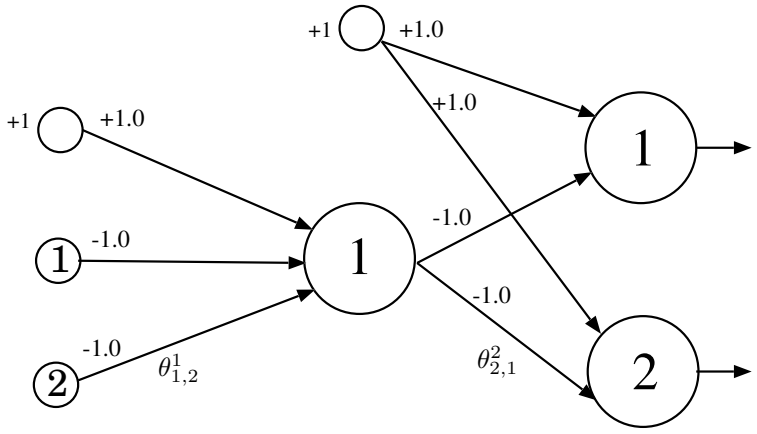


c) Clasificación de la muestra $(1, 3.5)^t$:

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:

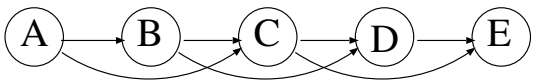
$$\begin{aligned}
&P(A = 1) \;=\; 0.3 \\
&P(B = 1 \mid A = 1) \;=\; 0.4 \\
&P(B = 1 \mid A = 0) \;=\; 0.6 \\
&P(C = 1 \mid A = 0, B = 0) \;=\; P(D = 1 \mid B = 0, C = 0) \;=\; P(E = 1 \mid C = 0, D = 0) \;=\; 0.2 \\
&P(C = 1 \mid A = 0, B = 1) \;=\; P(D = 1 \mid B = 0, C = 1) \;=\; P(E = 1 \mid C = 0,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

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?

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

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

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

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