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

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

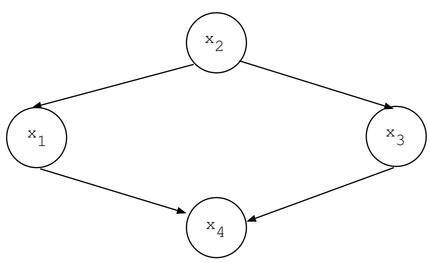
se cumplen las condiciones complementarias de Karush-Kuhn-Tucker $\alpha_i^\star v_i(\mathbf{\Theta}^\star) = 0$ para $1 \leq i \leq k$. Indicar cuál de las siguientes afirmaciones se deduce de ellas:

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

3 $\boxed{\text{C}}$ 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 $\boxed{\text{D}}$ 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:

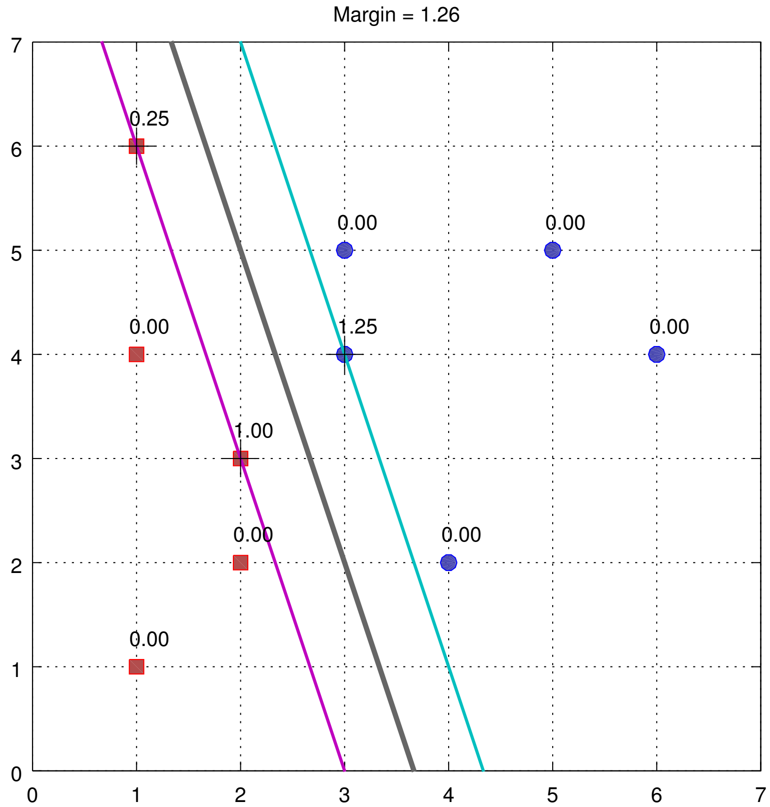
$$\begin{aligned}\theta_1^\star &= +1\cdot 0.25\cdot 1\ +1\cdot 1.0\cdot 2\ -1\cdot 1.25\cdot 3 = -1.5\\ \theta_2^\star &= +1\cdot 0.25\cdot 6\ +1\cdot 1.0\cdot 3\ -1\cdot 1.25\cdot 4 = -0.5\end{aligned}$$

- Peso umbral (con el 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:

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

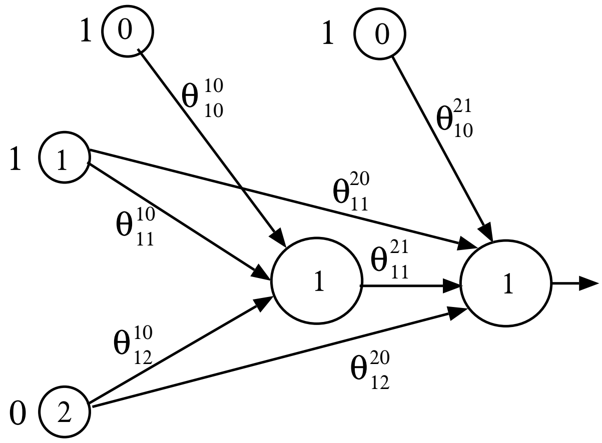
Alternativamente:

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

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 $\theta^{10}_{10} = \theta^{10}_{11} = \theta^{10}_{12} = \theta^{20}_{11} = \theta^{20}_{12} = \theta^{21}_{10} = \theta^{21}_{11} = 1.0$, 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

$\phi^1_1 = \theta^{10}_{10} + \theta^{10}_{11}\ x_1 + \theta^{10}_{12}\ x_2 = 2.0$ $\qquad s^1_1 = \dfrac{1}{1 + \exp(-\phi^1_1)} = .880797$

$\phi^2_1 = \theta^{21}_{10} + \theta^{20}_{11}\ x_1 + \theta^{20}_{10}\ x_2 + \theta^{21}_{11}\ s^1_1 = 2.880797$ $\qquad s^2_1 = \dfrac{1}{1 + \exp(-\phi^2_1)} = .946889$

b) El error en la capa de salida es:

$\delta^2_1 = (t_1 - s^2_1)\ s^2_1\ (1 - s^2_1) = .002671$

El error en la capa de oculta es:

$\delta^1_1 = (\delta^2_1\ \theta^{21}_{11})\ s^1_1\ (1 - s^1_1) = .000280$

c) Los nuevos pesos son:

$\theta^{21}_{10} = \theta^{21}_{10} + \rho\ \delta^2_1\ (+1) = \theta^{21}_{10} + 0.002671 = 1.002671$

$\theta^{21}_{11} = \theta^{21}_{11} + \rho\ \delta^2_1\ s^2_1 = \theta^{21}_{11} + 0.002529 = 1.002529$

$\theta^{20}_{11} = \theta^{20}_{11} + \rho\ \delta^2_1\ x_1 = \theta^{20}_{11} + 0.002671 = 1.002671$

$\theta^{20}_{12} = \theta^{20}_{12} + \rho\ \delta^2_1\ x_2 = \theta^{20}_{12} + 0.0 = 1.0$

$\theta^{10}_{10} = \theta^{10}_{10} + \rho\ \delta^1_1\ (+1) = \theta^{10}_{10} + 0.000280 = 1.000280$

$\theta^{10}_{11} = \theta^{10}_{11} + \rho\ \delta^1_1\ x_1 = \theta^{10}_{11} + 0.000280 = 1.000280$

$\theta^{10}_{12} = \theta^{10}_{12} + \rho\ \delta^1_1\ x_2 = \theta^{10}_{11} + 0.0 = 1.0$

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: $P(W = \mathrm{a}) = P(W = \mathrm{b}) = P(W = \mathrm{c})$,
- $P(X \mid W)$, $P(Y \mid X)$ 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

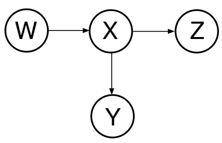
$P(z \mid x)$	z: a	b	c
x: a	1/3	0	2/3
b	1/4	1/2	1/4
c	0	3/5	2/5

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 $P(X=\mathrm{b},\ Y=\mathrm{b},\ Z=\mathrm{b} \mid W=\mathrm{b})$

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

a) Representación gráfica de la red:



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

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

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

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

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

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