

Aprendizaje Automático

Cuestiones y ejercicios

Francisco Casacuberta Nolla & Enrique Vidal Ruiz

Escola Tècnica Superior d'Informàtica
Dep. de Sistemes Informàtics i Computació
Universitat Politècnica de València

Julio, 2018

Tema 3: Técnicas de optimización

Cuestiones

- 1 [A] En la optimización directa de una función convexa q , $\nabla q(\Theta^*) = \mathbf{0}$ es una condición *necesaria y suficiente* para que Θ^* sea solución. ¿Y si no es convexa?
- A) $\nabla q(\Theta^*) = \mathbf{0}$ es una condición necesaria
 - B) $\nabla q(\Theta^*) = \mathbf{0}$ es una condición suficiente
 - C) $\nabla q(\Theta^*) = \mathbf{0}$ no es condición necesaria ni suficiente.
 - D) $\nabla q(\Theta^*) \neq \mathbf{0}$
- 2 [B] La técnica de descenso por gradiente garantiza, bajo determinadas condiciones, alcanzar un mínimo local. ¿De qué depende que sea un mínimo u otro?
- A) Del factor de aprendizaje
 - B) De la inicialización
 - C) De la forma de la función q a optimizar
 - D) Del número de mínimos locales.
- 3 [B] La técnica de descenso por gradiente garantiza, bajo determinadas condiciones, alcanzar un mínimo local. ¿En que caso ese mínimo es único?
- A) Nunca
 - B) La función q es convexa
 - C) La función q es polinómica
 - D) Siempre
- 4 [B] El algoritmo perceptrón cuando se aplica a una muestra de entrenamiento que no es linealmente separable verifica:
- A) Converge en un número de iteraciones doble al número de muestras
 - B) No converge nunca
 - C) A veces converge y a veces no
 - D) Converge en un número de iteraciones fijo

Problemas

1. Estimar los parámetros μ de una simple gaussiana multivariada (en \mathbb{R}^d) a partir de un conjunto de entrenamiento $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ y suponiendo que Σ está fijada a una matriz unidad:

$$p(\mathbf{x} \mid \Theta) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

Solución:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$$

2. Minimizar $q(\theta) = \theta_1^2 + \theta_2^2$ con $\theta_1 + \theta_2 = 1$

Solución:

$$\Lambda(\theta, \beta) = \theta_1^2 + \theta_2^2 + \beta(1 - \theta_1 - \theta_2)$$

$$\frac{\partial \Lambda}{\partial \theta_1} = 2\theta_1 - \beta = 0 \qquad \frac{\partial \Lambda}{\partial \theta_2} = 2\theta_2 - \beta = 0$$

$$\theta_1^*(\beta) = \frac{\beta}{2} \qquad \theta_2^*(\beta) = \frac{\beta}{2}$$

$$\Lambda_D(\beta) = \frac{\beta^2}{2} + \beta(1 - \beta) = -\frac{\beta^2}{2} + \beta$$

$$\frac{d\Lambda_D}{d\beta} = -\beta + 1 = 0 \Rightarrow \beta^* = 1$$

$$\theta_1^*(\beta^*) = \theta_2^*(\beta^*) = \frac{1}{2}$$

3. Aplicar la técnica de los mutiplicadores de Langrange al problema de la minimización de $q(\theta) = q_0 + (\theta - \theta_0)^2$ con $\theta \leq \theta_1$ en el caso de que $\theta_1 > \theta_0$

Solución: Aplicar la condición KKT.

4. Aplicar la técnica de los mutiplicadores de Langrange al problema de la minimización de $q(\theta) = q_0 + (\theta - \theta_0)^2$ con $q(\theta) + \theta = K$, donde K es una constante

5. Demostrar que en cualquier problema de clasificación en C clases, la estimación de máxima verosimilitud de la probabilidad a priori de cada clase c , $1 \leq c \leq C$, es $\hat{p}_c = n_c/N$, donde $N = \sum_c n_c$ es el número total de datos observados y n_c es el número de datos de la clase c .

Solución: Similar al visto en clase para dos clases.

6. Calcular los gradientes de las funciones perceptrón y de Widrow-Hoff

Solución: Gradiente de la función perceptrón

$$\begin{aligned}\nabla q_S(\theta) &= \nabla \sum_{\substack{(\mathbf{x},c) \in S \\ c \theta^t \mathbf{x} < 0}} -c \theta^t \mathbf{x} \\ &= \nabla \sum_{\substack{(\mathbf{x},c) \in S \\ c \theta^t \mathbf{x} < 0}} -c \sum_i \theta_i x_i \\ \frac{\partial q_S(\theta)}{\partial \theta_j} &= \sum_{\substack{(\mathbf{x},c) \in S \\ c \theta^t \mathbf{x} < 0}} -c x_j \\ \nabla q_S(\theta) &= \sum_{\substack{(\mathbf{x},c) \in S \\ c \theta^t \mathbf{x} < 0}} -c \mathbf{x}\end{aligned}$$

Gradiente de la función de Widrow -Hoff

$$\begin{aligned}\nabla q_S(\theta) &= \nabla \frac{1}{2} \sum_{(\mathbf{x},y) \in S} (\theta^t \mathbf{x} - y)^2 \\ &= \nabla \frac{1}{2} \sum_{(\mathbf{x},y) \in S} \left(\sum_i \theta_i x_i - y_i \right)^2 \\ \frac{\partial q_S(\theta)}{\partial \theta_j} &= \sum_{(\mathbf{x},y) \in S} \left(\sum_i \theta_i x_i - y_i \right) x_j \\ \nabla q_S(\theta) &= \sum_{(\mathbf{x},y) \in S} (\theta^t \mathbf{x} - y) \mathbf{x}\end{aligned}$$

Tema 4: Máquinas de vectores soporte

Cuestiones

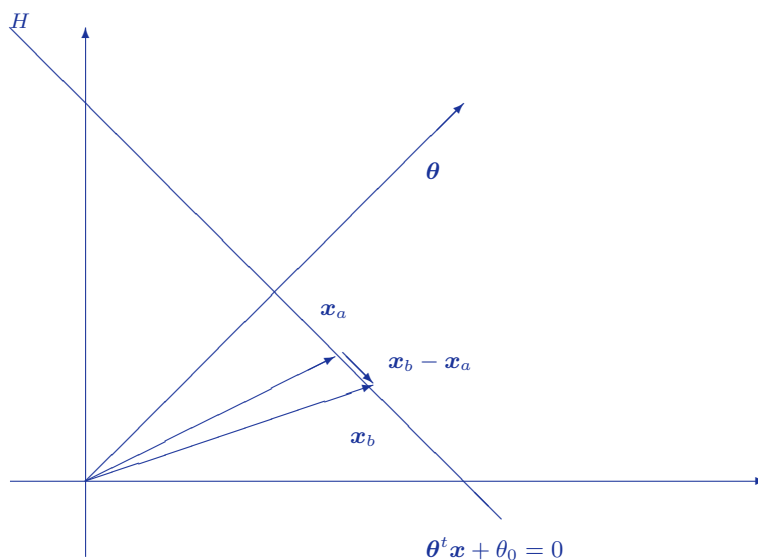
- 1 **B** En el caso de SVM para muestras linealmente separables, indicar qué afirmación es la incorrecta:
- A) Los vectores soporte están formados por al menos una muestra de cada clase.
 - B) Los vectores soporte están formados por al menos dos muestra de una clase y una muestra de la otra clase..
 - C) Los vectores soporte son las únicas muestras que contribuyen en la obtención del vector de pesos
 - D) El umbral se defina a partir de un vector soporte
- 2 **C** En el caso de SVM con márgenes blandos, las muestras que no son vectores soporte ($n : \alpha_n^* = 0$), verifican que
- A) $\zeta_n^* > 0$
 - B) $\zeta_n^* < 0$
 - C) $\zeta_n^* = 0$
 - D) $\zeta_n^* \neq 0$

Problemas

1. Demostrar que en el caso de una función discriminante para dos clases, el valor de la función en un punto es proporcional a la distancia al hiperplano separador.

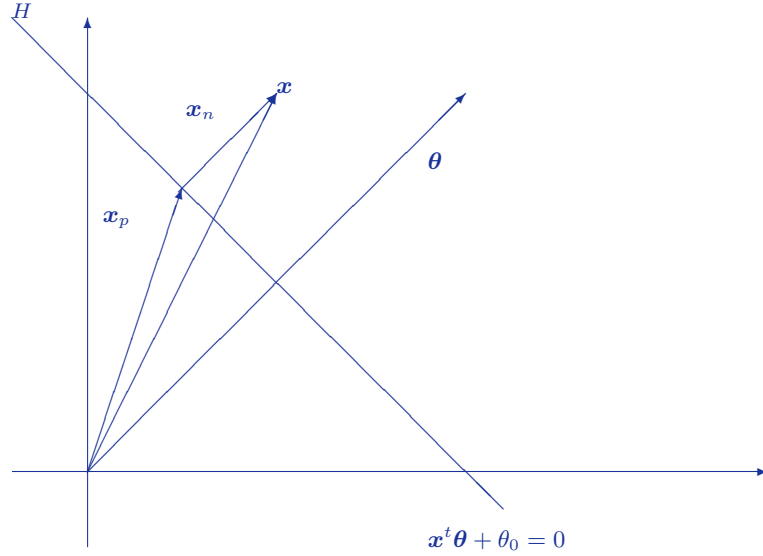
Solución:

- a) El vector de pesos es perpendicular al hiperplano separador



$$\begin{aligned} \mathbf{x}_a, \mathbf{x}_b \in H &\Rightarrow g(\mathbf{x}_a) = g(\mathbf{x}_b) = 0 \\ &\Rightarrow \boldsymbol{\theta}^t (\mathbf{x}_b - \mathbf{x}_a) = 0 \\ &\Rightarrow \boldsymbol{\theta} \perp H \end{aligned}$$

- b) El vector asociado a un punto cualquier se descompone en un vector al hiperplanos separador y un vector perpendicular al hiperplano separador



$$\begin{aligned}
 x = x_p + x_n &\Rightarrow x = x_p + r \frac{\theta}{\|\theta\|} \\
 &\Rightarrow g(x) = \theta_0 + \theta^t x \\
 &= \theta_0 + \theta^t x_p + r \frac{\theta^t \theta}{\|\theta\|} \\
 &= r \|\theta\|
 \end{aligned}$$

2. Dada una función discriminante lineal $\phi(x; \Theta)$, y un conjunto de puntos $S = \{(x_1, c_1), \dots, (x_N, c_N)\}$, encontrar la forma canónica con respecto a S del hiperplano que define la función discriminante lineal $\phi(x; \Theta)$.

Solución:

- a) $\hat{\phi} = \min_{1 \leq n \leq N} c_n (\theta^t x_n + \theta_0)$
b) $\theta_{ci} = \theta_i / \hat{\phi}$ para $0 \leq i \leq d$

3. Dada una función discriminante $\phi(x; \Theta)$, demostrar que si $\gamma \in \mathbb{R}^+$, entonces $\gamma \phi(x; \Theta)$ y $\phi(x; \Theta)$ representan la misma frontera de decisión.

Solución: Sean H' y H las fronteras de decisión asociadas a $\phi(x; \Theta)$ y a $\gamma \phi(x; \Theta)$, respectivamente. H' viene dado por la ecuación: $\phi(x; \Theta) = 0$; y H por: $\gamma \phi(x; \Theta) = 0$. Como $\gamma > 0$, la ecuación de H se puede reescribir como: $\phi(x; \Theta) = 0$, que es la misma ecuación que la de H' .

4. Desarrollar completamente los pasos necesarios hasta obtener la lagrangiana dual $\Lambda_D(\alpha)$. en el problema de la clasificación de margen máximo.

Solución:

$$\begin{aligned}
 \frac{\partial \Lambda(\theta, \theta_0, \alpha)}{\partial \theta_i} &= \theta_i - \sum_{n=1}^N c_n \alpha_n x_{ni} = 0 \text{ para } 1 \leq i \leq d \Rightarrow \theta^*(\alpha) = \sum_{n=1}^N c_n \alpha_n x_n \\
 \frac{\partial \Lambda(\theta, \theta_0, \alpha)}{\partial \theta_0} &= \sum_{n=1}^N c_n \alpha_n = 0
 \end{aligned}$$

$$\begin{aligned}
 \Lambda_D(\alpha) &= \Lambda(\theta^*, \theta_0^*, \alpha) \\
 &= \frac{1}{2} \theta^{*t} \theta^* - \sum_{n=1}^N \alpha_n (c_n (\theta^{*t} x_n + \theta_0^*) - 1) \\
 &= \frac{1}{2} \sum_{n, n'=1}^N c_n c_{n'} \alpha_n \alpha_{n'} x_n^t x_{n'} - \sum_{n=1}^N \alpha_n \left(c_n \left(\sum_{n'=1}^N c_{n'} \alpha_{n'} x_n^t x_{n'} + \theta_0^* \right) - 1 \right) \\
 &= -\frac{1}{2} \sum_{n, n'=1}^N c_n c_{n'} \alpha_n \alpha_{n'} x_n^t x_{n'} - \theta_0^* \sum_{n=1}^N \alpha_n c_n + \sum_{n=1}^N \alpha_n \\
 &= \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n, n'=1}^N c_n c_{n'} \alpha_n \alpha_{n'} x_n^t x_{n'}
 \end{aligned}$$

5. Sea $S = \{((1,1)^t, +1), ((2,2)^t, -1)\}$ una muestra de entrenamiento. Mediante el método de los multiplicadores de Lagrange, obtener (analíticamente) θ^* y θ^* que clasifiquen S con el máximo margen.

Solución: Pista: Resolver el problema *primal* y hacer uso de las condiciones KKT.

6. Sea $\phi(\mathbf{x}; \theta, \theta_0)$, $\mathbf{x}, \theta \in \mathbb{R}^2$, una FDL obtenida mediante el método SVM a partir de una muestra de entrenamiento en 2 dimensiones. Obtener las ecuaciones de: a) la recta de decisión asociada a ϕ ; b) las ecuaciones de las rectas que definen las fronteras del margen.

Solución:

a) Ecuación de la recta de separación asociada a ϕ :

$$\phi(\mathbf{x}; \theta, \theta_0) = 0 \Rightarrow \theta_1 x_1 + \theta_2 x_2 + \theta_0 = 0 \Rightarrow x_2 = -\frac{\theta_1}{\theta_2} x_1 - \frac{\theta_0}{\theta_2}$$

b) Ecuaciones de las rectas que definen las fronteras del margen:

$$\phi(\mathbf{x}; \theta, \theta_0) = +1 \Rightarrow \theta_1 x_1 + \theta_2 x_2 + \theta_0 = +1 \Rightarrow x_2 = -\frac{\theta_1}{\theta_2} x_1 - \frac{\theta_0 - 1}{\theta_2}$$

$$\phi(\mathbf{x}; \theta, \theta_0) = -1 \Rightarrow \theta_1 x_1 + \theta_2 x_2 + \theta_0 = -1 \Rightarrow x_2 = -\frac{\theta_1}{\theta_2} x_1 - \frac{\theta_0 + 1}{\theta_2}$$

7. Sea S una muestra linealmente separable. Demostrar que el margen óptimo es:

$$2 \left(\sum_{n \in \mathcal{V}} \alpha_n^* \right)^{-1/2}$$

Solución:

$$\begin{aligned} \|\theta^*\|^2 &= \theta^{*t} \theta^* \\ &= \sum_{n \in \mathcal{V}} c_n \alpha_n^* \theta^{*t} \mathbf{x}_n \\ &= \sum_{n \in \mathcal{V}} \alpha_n^* (1 - c_n \theta_0^*) \\ &= \sum_{n \in \mathcal{V}} \alpha_n^* - \theta_0^* \sum_{n \in \mathcal{V}} \alpha_n^* c_n \\ &= \sum_{n \in \mathcal{V}} \alpha_n^* \\ \frac{2}{\|\theta\|} &= \frac{2}{\sqrt{\sum_{n \in \mathcal{V}} \alpha_n^*}} \end{aligned}$$

8. Desarrollar completamente los pasos necesarios para encontrar una SVM para el caso que no es linealmente separable.

Solución: Los mismos pasos que en el caso linealmente separable r

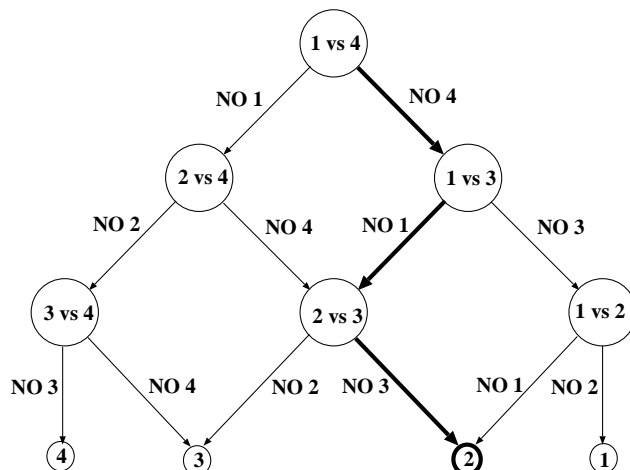
9. Linealizar una función discriminante cúbica

Solución:

Similar al desarrollo visto en clase con la función cuadrática.

10. Construir un DAG para un problema de clasificación en cuatro clases.

Solución:



11. Dado una muestra de entrenamiento linealmente separable

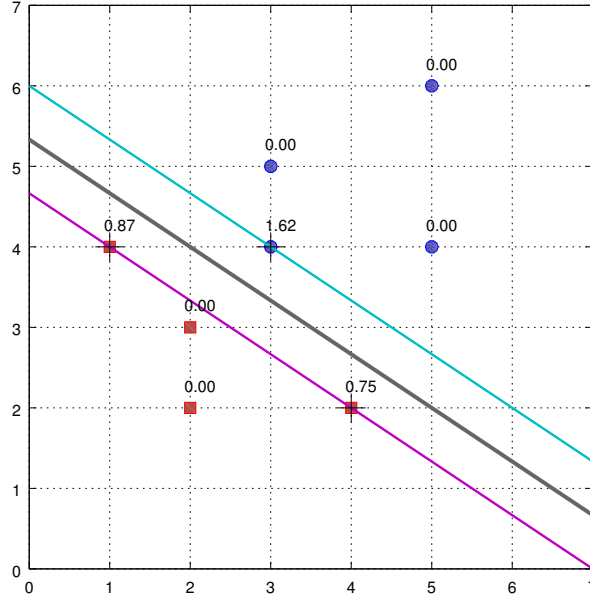
$$S = \{((1, 4), +1), ((2, 2), +1), ((2, 3), +1), ((4, 2), +1), ((3, 4), -1), ((3, 5), -1), ((5, 4), -1), ((5, 6), -1)\},$$

al calcular los multiplicadores de Lagrange óptimos α^* se obtiene $[0.87, 0.0, 0.0, 0.75, 1.62, 0.0, 0.0, 0.0]$. Obtener la correspondiente función discriminante lineal, calcular el margen óptimo y clasificar la muestra $(4, 5)$.

Solución:

- Vector de pesos y umbral:

$$\begin{aligned}\theta^* &= \sum_{n \in \mathcal{V}} c_n \alpha_n^* \mathbf{x}_n = 0.87 (1, 4)^t + 0.75 (4, 2)^t - 1.62 (3, 4)^t = (-1.0, -1.5)^t \\ \theta_0^* &= c_5 - \theta^{*t} \mathbf{x}_5 = 8.0\end{aligned}$$



- El margen óptimo es

$$2 \left(\sum_{n \in \mathcal{V}} \alpha_n^* \right)^{-1/2} = \frac{2}{\sqrt{0.87 + 0.75 + 1.62}} = 1.11$$

- La clasificación de $\mathbf{x} = (4, 5)$ será

$$(-1.0, -1.5)^t (4, 5) + 8.0 = -3.5 < 0 \Rightarrow \text{clase } -1$$

12. Dado una muestra de entrenamiento

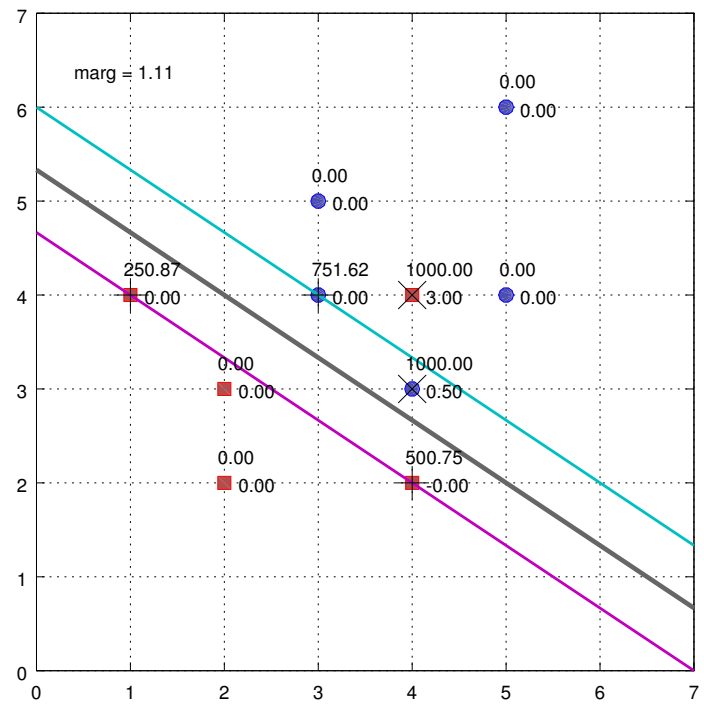
$$S = \{((1, 4), +1), ((2, 2), +1), ((2, 3), +1), ((4, 2), +1), ((3, 4), -1), ((3, 5), -1), ((5, 4), -1), ((5, 6), -1), ((4, 4), +1), ((4, 3), -1)\},$$

al calcular los multiplicadores de Lagrange óptimos α^* con $C = 1000$ se obtiene

$$[250.87, 0.0, 0.0, 500.75, 751.62, 0.0, 0.0, 0.0, 1000.0, 1000.0]$$

Obtener la correspondiente función discriminante lineal, las tolerancias óptimas ζ_n^* y clasificar la muestra $(4, 5)$.

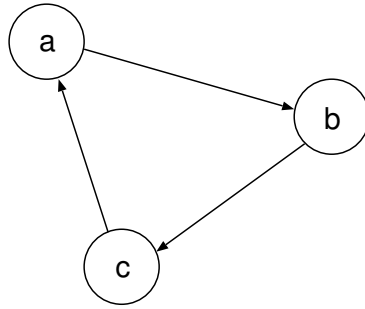
Solución: Procedimiento similar al problema anterior, Para calcular las tolerancias óptimas ζ_n^* , utilizar la condición KKT (1) de la transparencia “SVM en el caso de no separabilidad lineal” de los apuntes de APR. Los resultados se muestran en la siguiente figura.



Tema 5: Modelos gráficos

Cuestiones

1 [B](#) El grafo de la figura



representa una de las siguientes distribuciones conjuntas:

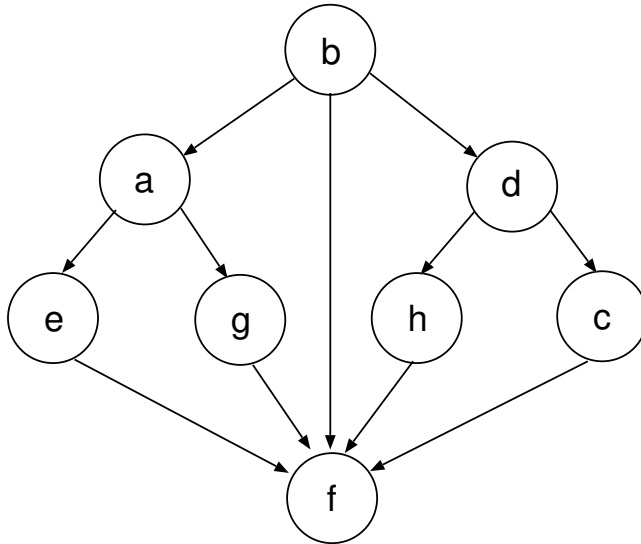
- A) $P(a, b, c) = P(a) P(b) P(c)$
- B) A ninguna distribución
- C) $P(a, b, c) = P(a | c) P(b | a) P(c | b)$
- D) $P(a, b, c) = P(a | c, b) P(b | a, c) P(c | b, a)$

Problemas

1. Dibujar la red bayesiana asociada a

$$P(a, b, c, d, e, f, g, h) = P(b) P(a | b) P(d | b) P(e | a) P(g | a) P(h | d) P(c | d) P(f | b, e, g, h, c)$$

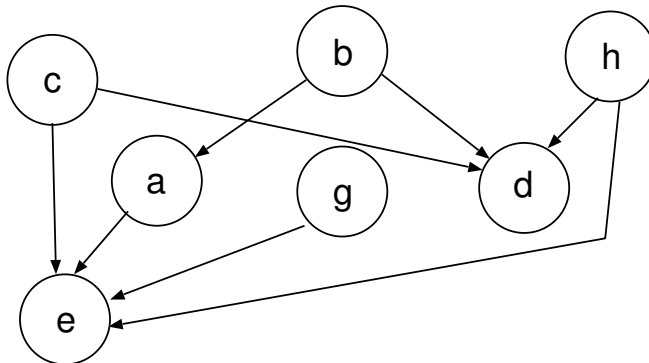
[Solución:](#)



2. Dibujar la red bayesiana asociada a

$$P(a, b, c, d, e, g, h) = P(b) P(c) P(h) P(a | b) P(g) P(d | c, b, h) P(e | c, a, g, h)$$

[Solución:](#)

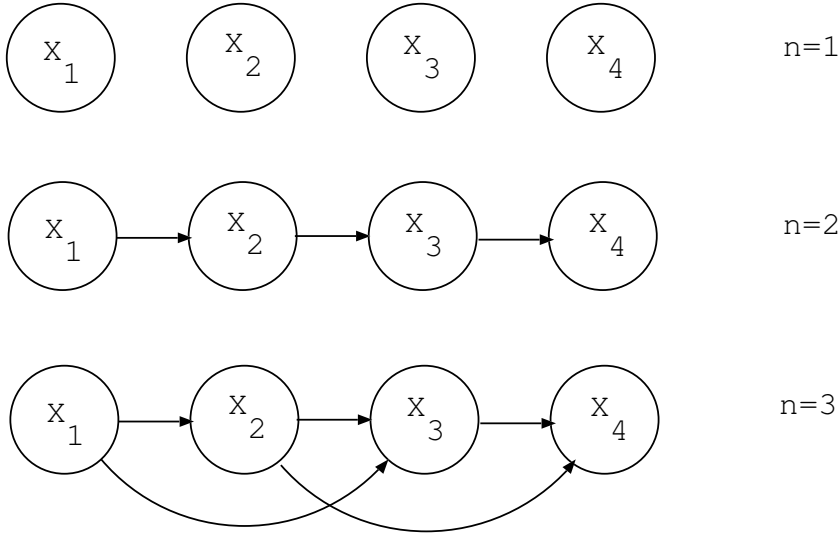


3. Las frases de un lenguaje se pueden modelar probabilísticamente mediante n -gramas, esto es la probabilidad de una palabra depende de las $n - 1$ últimas palabras:

$$\begin{aligned} P(X_1, X_2, \dots, X_N) &= P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2, X_3) \dots P(X_N | X_1, X_2, \dots, X_{N-1}) \\ &\approx P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2, X_3) \dots P(X_i | X_{i-n+2}, \dots, X_{i-1}) \dots \\ &\quad P(X_N | X_{N-n+2}, X_2, \dots, X_{N-1}) \end{aligned}$$

Para $N = 4$, construir las redes bayesianas en los casos en que $n = 1, 2, 3$

Solución:

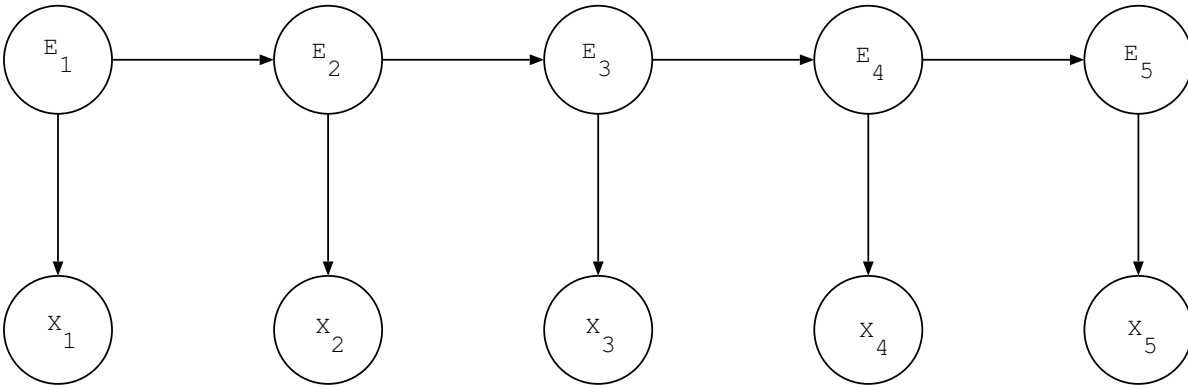


4. Dado un espacio de observaciones \mathbb{R}^d , un modelo de Markov oculto (ver asignatura SIN) $\mathcal{H} = (\mathcal{Q}, P_T, p_O, P_I)$, donde \mathcal{Q} es un conjunto finito de estado, $P_T(E' = e' | E = e)$ es la probabilidad de pasar del estado $e \in \mathcal{Q}$ al estado $e' \in \mathcal{Q}$, $P_O(X = \mathbf{x} | E = e)$ es el valor de la densidad de probabilidad de emitir $\mathbf{x} \in \mathbb{R}^d$ en estado $e \in \mathcal{Q}$ y $P_I(E = e)$ es la probabilidad de que el estado $e \in \mathcal{Q}$ sea estado inicial. La probabilidad de que \mathcal{H} genere una cadena dada de N observaciones $\mathbf{x}_1, \dots, \mathbf{x}_N$ con $\mathbf{x}_i \in \mathbb{R}^d$ para $1 \leq i \leq N$ es:

$$p_{\mathcal{H}}(X_1 = \mathbf{x}_1, \dots, X_N = \mathbf{x}_N) = \sum_{e_1, \dots, e_N \in \mathcal{Q}} P_I(E_1 = e_1) P_O(X_1 = \mathbf{x}_1 | E_1 = e_1) P_T(E_2 = e_2 | E_1 = e_1) P_O(X_2 = \mathbf{x}_2 | E_2 = e_2) \dots P_T(X_N = e_N | E_{N-1} = e_{N-1}) P_O(X_N = \mathbf{x}_N | E_N = e_N)$$

Se pide construir al red bayesiana que permita calcular $p_{\mathcal{H}}(X_1 = \mathbf{x}_1, \dots, X_5 = \mathbf{x}_5)$

Solución:



Esta red bayesiana permite representar la probabilidad conjunta

$$P(X_1 = \mathbf{x}_1, \dots, X_5 = \mathbf{x}_5, E_1 = e_1, \dots, E_5 = e_5) = P_I(E_1 = e_1) P_O(X_1 = \mathbf{x}_1 | E_1 = e_1) P_T(E_2 = e_2 | E_1 = e_1) P_O(X_2 = \mathbf{x}_2 | E_2 = e_2) \dots P_T(X_5 = e_5 | E_4 = e_4) P_O(X_5 = \mathbf{x}_5 | E_5 = e_5)$$

y $p_{\mathcal{H}}(X_1 = \mathbf{x}_1, \dots, X_5 = \mathbf{x}_5)$ se calcula como

$$p_{\mathcal{H}}(X_1 = \mathbf{x}_1, \dots, X_5 = \mathbf{x}_5) = \sum_{e_1, \dots, e_N \in \mathcal{Q}} P(X_1 = \mathbf{x}_1, \dots, X_5 = \mathbf{x}_5, E_1 = e_1, \dots, E_5 = e_5)$$

5. En el ejemplo sobre el depósito de combustible, batería e indicador eléctrico, comprobar que $P(C = 0 | I = 0, B = 0) \approx 0.111$

Solución: Desarrollo similar al que está en las transparencias

6. En la inferencia en cadenas, ¿Qué ocurre si también conocemos $x_{i'} \in E_{x_n}^+$ con $i' < i$? y ¿Qué ocurre si también conocemos $x_{f'} \in E_{x_n}^-$ con $f' > f$?

Solución:

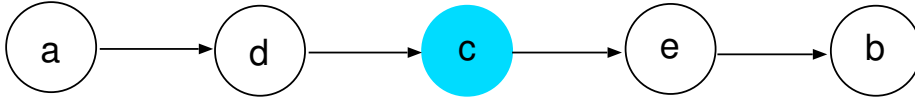
$$\begin{aligned} \blacksquare P(x_n | x_{i'}, x_i, x_f) &= \frac{P(x_n, x_{i'} | x_i, x_f)}{P(x_{i'} | x_i, x_f)} = \frac{P(x_n | x_i, x_f) P(x_{i'} | x_i, x_f)}{P(x_{i'} | x_i, x_f)} = P(x_n | x_i, x_f) \\ \blacksquare P(x_n | x_i, x_f, x_{f'}) &= P(x_n | x_i, x_f) \end{aligned}$$

7. En la inferencia en cadenas, calcular $P(x_n)$

Solución: Si la cadena es de longitud N

$$\begin{aligned} \pi(x_1) &= P(x_1) & \lambda(x_N) &= P(x_N | x_{N-1}) \\ \pi(x_n) &= \sum_{x_{n-1}} P(x_n | x_{n-1}) \pi(x_{n-1}) & \lambda(x_n) &= \sum_{x_{n+1}} P(x_{n+1} | x_n) \lambda(x_{n+1}) \\ P(x_n) &= \alpha \pi(x_n) \lambda(x_n) \end{aligned}$$

8. Dada la siguiente red bayesiana



Demostrar que $P(a, b | c) = P(a | c) P(b | c)$

Solución:

$$\begin{aligned} P(a, b | c) &= \frac{P(a, c, b)}{P(c)} \\ &= \frac{\sum_{d, e} P(a, d, c, e, b)}{P(c)} \\ &= \frac{\sum_{d, e} P(a) P(d | a) P(c | d) P(e | c) P(b | e)}{P(c)} \\ &= \frac{P(a) \sum_d (P(d | a) P(c | d)) \sum_e (P(e | c) P(b | e))}{P(c)} \\ &= \frac{P(a) \sum_d P(d, c | a) \sum_e P(b, e | c)}{P(c)} \\ &= \frac{P(a) P(c | a) P(b | c)}{P(c)} \\ &= P(a | c) P(b | c) \end{aligned}$$

9. Considerar la red bayesiana \mathcal{R} definida como $P(U, V, X, Y) = P(U) P(X | U) P(V | U) P(Y | V)$, cuyas variables U, V , toman valores en el conjunto $\{1, 2\}$ y las variables X, Y en el conjunto $\{\text{"a"}, \text{"b"}\}$ y las distribuciones de probabilidad asociadas son como sigue:

- $P(U)$ viene dada por $P(U = 1) = 2/5, P(U = 2) = 3/5$
- $P(V | U)$ viene dada por la tabla A
- $P(X | U)$ y $P(Y | V)$ son idénticas y vienen dadas por la tabla B

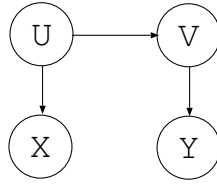
A	1	2
1	0	1
2	1	0

B	"a"	"b"
1	3/4	1/4
2	1/3	2/3

- a) Representar gráficamente la red
- b) Obtener una expresión simplificada de $P(X, Y)$ en función de las distribuciones que definen \mathcal{R} y calcular $P(X = \text{"a"}, Y = \text{"a"})$
- c) ¿Cuáles son los valores de u, v para los que $P(U, V, X = \text{"a"}, Y = \text{"a"})$ es máxima?
- d) \mathcal{R} corresponde a un proceso de generación de una cadena de dos símbolos mediante un modelo de Markov. Representar gráficamente este modelo.

Solución:

a) Representación gráfica de la red:



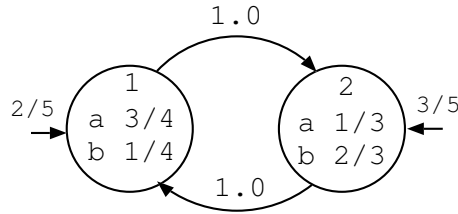
b) Obtener una expresión simplificada de $P(X, Y)$ en función de las distribuciones que definen \mathcal{R} :

$$\begin{aligned}
 P(X, Y) &= \sum_u \sum_v P(U = u) P(X | U = u) P(V = v | U = u) P(Y | V = v) \\
 &= \sum_u P(U = u) P(X | U = u) \sum_v P(V = v | U = u) P(Y | V = v) \\
 &\equiv \sum_u P(u) P(X | u) \sum_v P(v | u) P(Y | v) \\
 P(X = "a", Y = "a") &= \sum_u P(u) P("a" | u) \sum_v P(v | u) P("a" | v) \\
 &= \sum_u P(u) P("a" | u) \left(P(V = 1 | u) P("a" | V = 1) + P(V = 2 | u) P("a" | V = 2) \right) \\
 &= \sum_u P(u) P("a" | u) \left(\frac{3}{4} P(V = 1 | u) + \frac{1}{3} P(V = 2 | u) \right) \\
 &= \left(\frac{3}{5} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot 0 + \frac{3}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} \cdot 1 + \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{3}{4} \cdot 1 + \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot 0 \right) = \frac{3}{20} + \frac{2}{20} = \frac{1}{4}
 \end{aligned}$$

c) Valores de U, V para los que $P(U, V, X = "a", Y = "a")$ es máxima:

$$\begin{aligned}
 \hat{u}, \hat{v} &= \arg \max_{u, v} P(U = u, V = v | X = "a", Y = "a") \\
 &= \arg \max_{u, v} \frac{P(U = u, V = v, X = "a", Y = "a")}{P(X = "a", Y = "a")} = \arg \max_{u, v} P(U = u, V = v, X = "a", Y = "a") \\
 \dots &\text{ Solo hay dos combinaciones no nulas: } U = 1, V = 2 \text{ y } U = 2, V = 1. \\
 &\text{De estas la máxima se obtiene con: } \hat{u} = 1, \hat{v} = 2
 \end{aligned}$$

d) Modelo de Markov representado por \mathcal{R} :



Tema 6: Redes neuronales multicapa

Cuestiones

- 1 [C] La derivada de la función sigmoid $g'_S(z) = \frac{d g_S}{d z}$ verifica una de las siguientes propiedades
- A) $g'_S(z) = g_S^2(z)$
 - B) $g'_S(z) = g_S(z)(1 - g_S(z))^2$
 - C) $g'_S(z) = g_S(z)(1 - g_S(z))$
 - D) $g'_S(z) = (1 - g_S(z))$
- 2 [A] La derivada de la función tangente hiperbólica $g'_T(z) = \frac{d g_T}{d z}$ verifica una de las siguientes propiedades
- A) $g'_T(z) = 1 - (g_T(z_k))^2$
 - B) $g'_T(z) = 1 - g_T(z_k)$
 - C) $g'_T(z) = (1 - g_T(z_k)) g_T(z_k)$
 - D) $g'_T(z) = (g_T(z_k))^2$
- 3 [B] La parálisis de una red multicapa se produce cuando:
- A) En entrenamiento se alcanza un mínimo de la función de error.
 - B) En entrenamiento, la red no evoluciona porque la derivada de la función de activación es muy pequeña
 - C) En clasificación, la red devuelve un cero
 - D) En entrenamiento, se acaban las muestras de entrenamiento

Problemas

1. Demostrar que $g_T(z) = 2g_S(2z) - 1 \quad \forall z \in \mathbb{R}$
2. Demostrar que la derivada de la función Softmax es, dados $z_1, \dots, z_n \in \mathbb{R}$,
$$g_M(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \Rightarrow g'_M(z_i) = \frac{d g_M}{d z} = g_M(z_i) (1 - g_M(z_i))$$
3. Supongamos que queremos resolver un problema de clasificación en 3 clases con un perceptrón multicapa con dos capas ocultas ($M_3 = 3$). Se emplean funciones de activación en escalón en las capas ocultas y softmax en la capa de salida. Las muestras se representan en un espacio de representación de dos dimensiones ($M_0 = 2$). Calcula la salida de la red neuronal para la muestra $\mathbf{x} = (-1, 2)$ con los siguientes vectores de pesos:

$$\begin{aligned}\theta_1^1 &= [2, -2, 0] & \theta_2^1 &= [2, -1, -1] & \theta_3^1 &= [-2, 1, 1] \\ \theta_1^2 &= [0, 1, 0, -1] & \theta_2^2 &= [0, 2, -3, 0] \\ \theta_1^3 &= [1, 1, 1] & \theta_2^3 &= [-1, 0, 1] & \theta_3^3 &= [1, 2, 1]\end{aligned}$$

Solución:

- Salida de la primera capa oculta

$$\begin{aligned}\phi_1^1 &= \theta_1^{1t} \mathbf{x} = 2 \cdot 1 + (-2) \cdot (-1) + 0 \cdot 2 = 4 & s_1^1 &= g(\phi_1^1) = 1 \\ \phi_2^1 &= \theta_2^{1t} \mathbf{x} = 2 \cdot 1 + (-1) \cdot (-1) + (-1) \cdot 2 = 1 & s_2^1 &= g(\phi_2^1) = 1 \\ \phi_3^1 &= \theta_3^{1t} \mathbf{x} = (-2) \cdot 1 + 1 \cdot (-1) + 1 \cdot 2 = -1 & s_3^1 &= g(\phi_3^1) = 0\end{aligned}$$

- Salida de la segunda capa oculta

$$\begin{aligned}\phi_1^2 &= \theta_1^{2t} \mathbf{s}_1 = 0 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 + (-1) \cdot 0 = 1 & s_1^2 &= g(\phi_1^2) = 1 \\ \phi_2^2 &= \theta_2^{2t} \mathbf{s}_1 = 0 \cdot 1 + 2 \cdot 1 + (-3) \cdot 1 + 0 \cdot 0 = -1 & s_2^2 &= g(\phi_2^2) = 0\end{aligned}$$

- Salida de la capa salida

$$\begin{aligned}\phi_1^3 &= \theta_1^{3t} \mathbf{s}_2 = 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 = 2 & \exp(\phi_1^3) &= 7.3891 \\ \phi_2^3 &= \theta_2^{3t} \mathbf{s}_2 = (-1) \cdot 1 + 0 \cdot 1 + 1 \cdot 0 = -1 & \exp(\phi_2^3) &= 0.36788 \\ \phi_3^3 &= \theta_3^{3t} \mathbf{s}_2 = 1 \cdot 1 + 2 \cdot 1 + 1 \cdot 0 = 3 & \exp(\phi_3^3) &= 20.086 \\ \exp(\phi_1^3) + \exp(\phi_2^3) + \exp(\phi_3^3) &= 27.843 \\ s_1^3 &= \frac{7.389}{27.843} = 0.26539 & s_2^3 &= \frac{0.36788}{27.843} = 0.013213 & s_3^3 &= \frac{20.086}{27.843} = 0.72140\end{aligned}$$

■ $\mathbf{x} = (-1, 2)$ es de la clase 3

4. Dado una muestra de entrenamiento $S = \{((1, 4), A), ((2, 3), B), ((4, 2), B), ((3, 5), C), ((5, 4), C)\}$, generar un conjunto de entrenamiento con los rangos de entrada normalizados.

Solución:

a) $S' = \{(1, 4), (2, 3), (4, 2), (3, 5), (5, 4)\}$

b) $\mu_1 = \frac{1 + 2 + 4 + 3 + 5}{5} = 3$

$$\mu_2 = \frac{4 + 3 + 2 + 5 + 4}{5} = 3.6$$

$$\sigma_1 = \frac{(-2)^2 + (-1)^2 + 1^2 + 0^2 + 2^2}{5 - 1} = 1.5811$$

$$\sigma_2 = \frac{1.6^2 + (-0.6)^2 + (-1.6)^2 + 1.4^2 + 0.4^2}{5 - 1} = 1.3784$$

c) Conversión:

$$(1, 4) \Rightarrow \left(\frac{1 - 3}{1.5811}, \frac{4 - 3.6}{1.3784} \right) = (-1.264, 0.29019)$$

$$(2, 3) \Rightarrow \left(\frac{2 - 3}{1.5811}, \frac{3 - 3.6}{1.3784} \right) = (-0.63247, -0.43529)$$

$$(4, 2) \Rightarrow \left(\frac{4 - 3}{1.5811}, \frac{2 - 3.6}{1.3784} \right) = (0.63247, -1.1608)$$

$$(3, 5) \Rightarrow \left(\frac{3 - 3}{1.5811}, \frac{5 - 3.6}{1.3784} \right) = (0.0, 1.0157)$$

$$(5, 4) \Rightarrow \left(\frac{5 - 3}{1.5811}, \frac{4 - 3.6}{1.3784} \right) = (1.2649, 0.29019)$$

5. Considerar un perceptrón de dos capas con la siguiente topología: dos entradas, tres nodos en la capa de salida y dos en la oculta, todos los nodos con función de activación sigmoid. Sus pesos son:

$$\begin{aligned}\boldsymbol{\theta}_1^1 &= (1, 1, 1)^t & \boldsymbol{\theta}_2^1 &= (1, 2, 1)^t \\ \boldsymbol{\theta}_1^2 &= (1, 1, 1)^t & \boldsymbol{\theta}_2^2 &= (-1, -2, -1)^t & \boldsymbol{\theta}_3^2 &= (-1, 2, -1)^t\end{aligned}$$

Detallar la traza de una iteración del algoritmo BackProp, con factor de aprendizaje $\rho = 1.0$, para una muestra de entrenamiento (\mathbf{x}, \mathbf{t}) , $\mathbf{x} = (-2, 1)^t$, $\mathbf{t} = (0, 1, 0)^t$.

Solución: Utilizaremos la notación extendida para \mathbf{x} con $x_0 = 1.0$: $\mathbf{x} = (-2, 1)^t \rightarrow \mathbf{x} = (1, -2, 1)^t$

a) Cálculo hacia adelante de las salidas de cada nodo:

$$\begin{aligned}1) \text{ Capa oculta: } \phi_1^1(\mathbf{x}) &= \boldsymbol{\theta}_1^1 \mathbf{x} = 0; & s_1^1 &= g(\phi_1^1(\mathbf{x})) = \frac{1}{1 + \exp(0)} = 0.5 \\ \phi_2^1(\mathbf{x}) &= \boldsymbol{\theta}_2^1 \mathbf{x} = -2; & s_2^1 &= g(\phi_2^1(\mathbf{x})) = \frac{1}{1 + \exp(2)} = 0.11920 \\ (\text{Notacion extendida}) & & s_0^1 &= 1.0 \\ 2) \text{ Capa de salida: } \phi_1^2(\mathbf{s}^1) &= \boldsymbol{\theta}_1^2 \mathbf{s}^1 = 1.61920; & s_1^2 &= g(\phi_1^2(\mathbf{s}^1)) = \frac{1}{1 + \exp(-1.61920)} = 0.83468 \\ \phi_2^2(\mathbf{s}^1) &= \boldsymbol{\theta}_2^2 \mathbf{s}^1 = -2.11920; & s_2^2 &= g(\phi_2^2(\mathbf{s}^1)) = \frac{1}{1 + \exp(2.11920)} = 0.10724 \\ \phi_3^2(\mathbf{s}^1) &= \boldsymbol{\theta}_3^2 \mathbf{s}^1 = -0.11920; & s_3^2 &= g(\phi_3^2(\mathbf{s}^1)) = \frac{1}{1 + \exp(0.11920)} = 0.47024\end{aligned}$$

b) Cálculo de errores hacia atrás:

1) Errores en la capa de salida:

$$\begin{aligned}\delta_1^2 &= (t_1 - s_1^2) g'(\phi_1^2) = (t_1 - s_1^2) s_1^2 (1 - s_1^2) = -0.11518 \\ \delta_2^2 &= (t_2 - s_2^2) g'(\phi_2^2) = (t_2 - s_2^2) s_2^2 (1 - s_2^2) = 0.08547 \\ \delta_3^2 &= (t_3 - s_3^2) g'(\phi_3^2) = (t_3 - s_3^2) s_3^2 (1 - s_3^2) = -0.11714\end{aligned}$$

2) Errores en la capa oculta:

$$\begin{aligned}\delta_1^1 &= (\delta_1^2 \theta_{11}^2 + \delta_2^2 \theta_{21}^2 + \delta_3^2 \theta_{31}^2) g'(\phi_1^1) = (\delta_1^2 \theta_{11}^2 + \delta_2^2 \theta_{21}^2 + \delta_3^2 \theta_{31}^2) s_1^1 (1 - s_1^1) = -0.12582 \\ \delta_2^1 &= (\delta_1^2 \theta_{12}^2 + \delta_2^2 \theta_{22}^2 + \delta_3^2 \theta_{32}^2) g'(\phi_2^1) = (\delta_1^2 \theta_{12}^2 + \delta_2^2 \theta_{22}^2 + \delta_3^2 \theta_{32}^2) s_2^1 (1 - s_2^1) = -0.00834\end{aligned}$$

c) Actualización de los pesos:

1) Actualización de pesos de la capa de salida:

$$\begin{aligned}\theta_{10}^2 &= \theta_{10}^2 + \Delta \theta_{10}^2 = \theta_{10}^2 + \rho \delta_1^2 s_0^1 = 0.88483 \\ \theta_{11}^2 &= \theta_{11}^2 + \Delta \theta_{11}^2 = \theta_{11}^2 + \rho \delta_1^2 s_1^1 = 0.94241 \\ \theta_{12}^2 &= \theta_{12}^2 + \Delta \theta_{12}^2 = \theta_{12}^2 + \rho \delta_1^2 s_2^1 = 0.98627 \\ \theta_{20}^2 &= \theta_{20}^2 + \Delta \theta_{20}^2 = \theta_{20}^2 + \rho \delta_2^2 s_0^1 = -0.91452 \\ \theta_{21}^2 &= \theta_{21}^2 + \Delta \theta_{21}^2 = \theta_{21}^2 + \rho \delta_2^2 s_1^1 = -1.95726 \\ \theta_{22}^2 &= \theta_{22}^2 + \Delta \theta_{22}^2 = \theta_{22}^2 + \rho \delta_2^2 s_2^1 = -0.98981 \\ \theta_{30}^2 &= \theta_{30}^2 + \Delta \theta_{30}^2 = \theta_{30}^2 + \rho \delta_3^2 s_0^1 = -1.11714 \\ \theta_{31}^2 &= \theta_{31}^2 + \Delta \theta_{31}^2 = \theta_{31}^2 + \rho \delta_3^2 s_1^1 = 1.94143 \\ \theta_{32}^2 &= \theta_{32}^2 + \Delta \theta_{32}^2 = \theta_{32}^2 + \rho \delta_3^2 s_2^1 = -1.01396\end{aligned}$$

2) Actualización de pesos de la capa oculta:

$$\begin{aligned}\theta_{10}^1 &= \theta_{10}^1 + \Delta \theta_{10}^1 = \theta_{10}^1 + \rho \delta_1^1 x_0 = 0.87418 \\ \theta_{11}^1 &= \theta_{11}^1 + \Delta \theta_{11}^1 = \theta_{11}^1 + \rho \delta_1^1 x_1 = 1.25163 \\ \theta_{12}^1 &= \theta_{12}^1 + \Delta \theta_{12}^1 = \theta_{12}^1 + \rho \delta_1^1 x_2 = 0.87418 \\ \theta_{20}^1 &= \theta_{20}^1 + \Delta \theta_{20}^1 = \theta_{20}^1 + \rho \delta_2^1 x_0 = 0.99166 \\ \theta_{21}^1 &= \theta_{21}^1 + \Delta \theta_{21}^1 = \theta_{21}^1 + \rho \delta_2^1 x_1 = 2.01668 \\ \theta_{22}^1 &= \theta_{22}^1 + \Delta \theta_{22}^1 = \theta_{22}^1 + \rho \delta_2^1 x_2 = 0.99166\end{aligned}$$