Examen de Aprendizaje Automático

ETSINF, Universitat Politècnica de València, 10 de enero de 2022

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

Cuestiones (2 puntos; tiempo estimado: 30 minutos)

Marca cada recuadro con una única opción de entre las dadas. Cada acierto suma 1/2 puntos y cada fallo resta 1/6 puntos.

1 $\boxed{\text{A}}$ 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 10000 muestras. Se han obtenido un total de 50 errores. Indicar cuál de las afirmaciones siguientes es razonable:

A) La talla de entrenamiento efectiva es 9900 muestras y el error estimado es $0.50\,\% \pm 0.14\,\%$
B) La talla de entrenamiento efectiva es de 9900 muestras y el error estimado es $5.0\,\%$
C) La talla de test efectiva es de 10000 muestras y el error estimado es $5.0\,\% \pm 0.14\,\%$
D) El error estimado es $0.50\,\% \pm 0.0014\,\%$.

2 $\boxed{\text{C}}$ Considerar la siguiente modificación de la función de Widrow y Hoff

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

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 esta expresión, el gradiente, $\boldsymbol{\nabla} q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)}$, es:

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

3 $\boxed{\text{D}}$ En el compromiso entre sesgo y varianza indicar qué afirmaión es correcta.

A) Si el modelo presenta una gran varianza quiere decir que el error de entrenamiento es alto.
B) Cuando el sesgo es muy alto, la varianza suele ser también alta.
C) Si el modelo presenta una sesgo alto quiere decir que el error de entrenamiento es bajo.
D) Cuando el sesgo es muy bajo, la varianza suele ser alta.

4 $\boxed{\text{D}}$ En el aprendizaje de modelos probabilísticos mediante el algoritmo esperanza-maximización indicar qué afirmaión es correcta.

A) Es el método adecuado cuando las muestra de entrenamiento es completa.
B) A partir de una cierta estimación de las variables ocultas se busca el óptimo de una función auxiliar como solución final del problema original.
C) Es el método adecuado cuando no hay variables ocultas.
D) En cada iteración, a partir de una cierta estimación de las variables ocultas se busca el óptimo de una función auxiliar.

Problema 1 (3 puntos; tiempo estimado: 30 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	2	2	2	3	4	3	1
x_{i2}	2	3	4	1	2	2	1	4
Clase	+1	-1	+1	–1	–1	–1	–1	+1
$\alpha_i^\star$	10.0	10.0	3.78	3.11	0.67	0	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 $(1, 2)^t$.

a) Pesos de la función discriminante:

$$\theta_1^* = (+1)\ (2)\ (10.0) + (-1)\ (2)\ (10.0) + (+1)\ (2)\ (3.79) + (-1)\ (2)\ (3.11) + (-1)\ (3)\ (0.67) = \ -0.67$$
$$\theta_2^* = (+1)\ (2)\ (10.0) + (-1)\ (3)\ (10.0) + (+1)\ (4)\ (3.79) + (-1)\ (1)\ (3.11) + (-1)\ (2)\ (0.67) = \ 0.67$$

Usando el vector soporte $\mathbf{x_4}$ (que verifica la condición : $0 < \alpha_1^* < C$)

$\theta_0^\star = c_4 - \boldsymbol{\theta}^{\star t}\mathbf{x_4} = 1 - ((-0.67)\ (2) + (0.67)\ (1)) = -0.33$

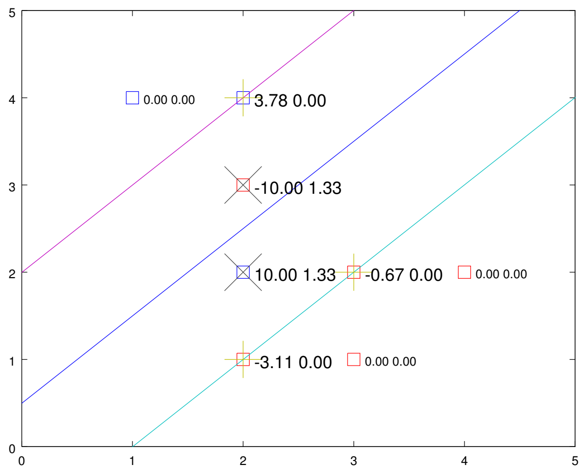
Función discriminante lineal: $\phi(\mathbf{x}) = -0.33 - 0.67\ x_1 + 0.67\ x_2$

b) Frontera de separación y representación gráfica:

Ecuación de la frontera lineal de separación: $-0.33 - 0.67\ x_1 + 0.67\ x_2 \ = \ 0 \quad \rightarrow \quad x_2 = 1.0\ x_1 + 0.49\,.$

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 ($i \in \{3, 4, 5, 6, 7, 8\}$) tienen una tolerancia $\zeta_i^* = 0$ y el resto

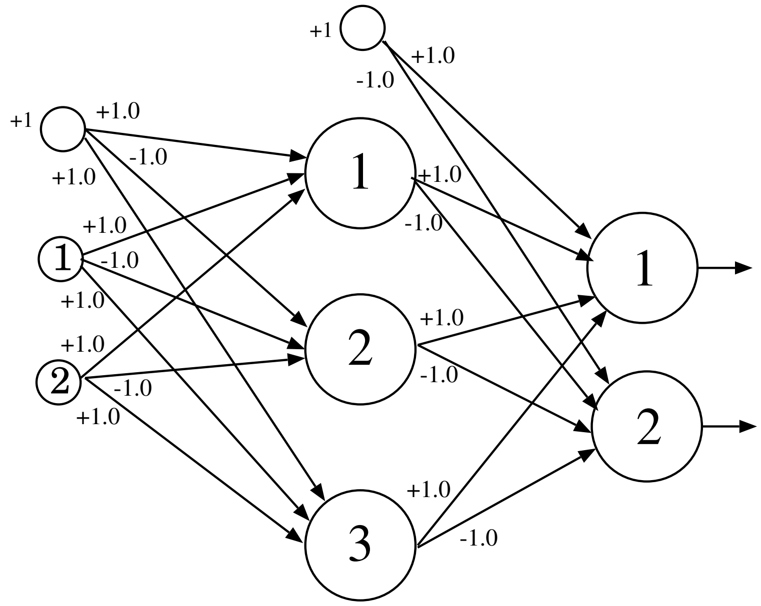
$\zeta_1^* \ = \ 1 - c_1\ (\boldsymbol{\theta}^{*t}\ \boldsymbol{x}_1 + \theta_0^*) \ = \ 1.33; \qquad \zeta_2^* \ = \ 1 - c_2\ (\boldsymbol{\theta}^{*t}\ \boldsymbol{x}_2 + \theta_0^*) \ = \ 1.33$

d) Clasificación de la muestra $(1, 2)^t$:

El valor de la función discriminante para este vector es: $\theta_0^* + 1\,\theta_1^* + 2\,\theta_2^* \ = \ 0.34 \ > \ 0 \ \Rightarrow$ 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 sigmoid, y factor de aprendizaje $\rho = 1.0$.



Dado un vector de entrada $x = (+1, +1)$ y su valor deseado de salida $t = (+1, 0)$,
Calcular:

a) las salidas que faltan

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

c) Los nuevos valores de los pesos de las conexiones $\theta^1_{3,2}$, que va del nodo 2 de entrada al nodo 3 de la capa oculta, y $\theta^2_{2,3}$, que va del nodo 3 de la capa oculta al nodo 2 de la capa de salida.

a) Las salidas de todos los nodos

$$\begin{aligned}
\phi^1_1 &= \theta^1_{1,1}\ x_1 + \theta^1_{1,2}\ x_2 + \theta^1_{1,0} = +3.0\,; & s^1_1 &= \tfrac{1}{1+\exp(-\phi^1_1)} = 0.953\\
\phi^1_2 &= \theta^1_{2,1}\ x_1 + \theta^1_{2,2}\ x_2 + \theta^1_{3,0} = -3.0\,; & s^1_2 &= \tfrac{1}{1+\exp(-\phi^1_2)} = 0.047\\
\phi^1_3 &= \theta^1_{3,1}\ x_1 + \theta^1_{3,2}\ x_2 + \theta^1_{3,0} = +3.0\,; & s^1_3 &= \tfrac{1}{1+\exp(-\phi^1_3)} = 0.953
\end{aligned}$$

$$\begin{aligned}
\phi^2_1 &= \theta^2_{1,1}\ s^1_1 + \theta^2_{1,2}\ s^1_2 + \theta^2_{1,3}\ s^1_3 + \theta^2_{1,0} = +2.953\,; & s^2_1 &= \tfrac{1}{1+\exp(-\phi^2_1)} = 0.950\\
\phi^2_2 &= \theta^2_{2,1}\ s^1_1 + \theta^2_{2,2}\ s^1_2 + \theta^2_{2,3}\ s^1_3 + \theta^2_{2,0} = -2.953\,; & s^2_2 &= \tfrac{1}{1+\exp(-\phi^2_2)} = 0.050
\end{aligned}$$

b) los correspondientes errores en los nodos de la capa de salida y en los nodos de la capa oculta:

$$\begin{aligned}
\delta^2_1 &= (t_1 - s^2_1)\ s^2_1\ (1 - s^2_1) &&= +0.0023\\
\delta^2_2 &= (t_2 - s^2_2)\ s^2_2\ (1 - s^2_2) &&= -0.0023
\end{aligned}$$

$$\begin{aligned}
\delta^1_1 &= (\delta^2_1\ \theta^2_{1,1} + \delta^2_2\ \theta^2_{2,1})\ s^1_1\ (1 - s^1_1) &&= +0.0002\\
\delta^1_2 &= (\delta^2_1\ \theta^2_{1,2} + \delta^2_2\ \theta^2_{2,2})\ s^1_2\ (1 - s^1_2) &&= +0.0002\\
\delta^1_3 &= (\delta^2_1\ \theta^2_{1,3} + \delta^2_2\ \theta^2_{2,3})\ s^1_3\ (1 - s^1_3) &&= +0.0002
\end{aligned}$$

b) El nuevo peso $\theta^2_{2,3}$ es: $\theta^2_{2,3} = \theta^2_{2,3} + \rho\ \delta^2_2\ s^1_3 = (-1.0) + (1)\ (-0.0023)\ (0.953) = -1.0022$

El nuevo peso $\theta^1_{3,2}$ es: $\theta^1_{3,2} = \theta^1_{3,2} + \rho\ \delta^1_3\ x_2 = (+1.0) + (1)\ (0.0002)\ (+1.0) = +1.0002$

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

Considerar la red bayesiana $\mathcal{R}$ definida como $P(R,T,X,Y,Z) = P(R)\ P(X \mid R)\ P(Y \mid R)\ P(Z \mid R)\ P(T \mid Y)$, cuyas variables R y T toman valores en $\{1,2,3\}$ y las variables X, Y, Z, en el conjunto $\{\texttt{"a"},\texttt{"b"}\}$. 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 A.
- $P(T \mid Y)$ viene dada por la tabla B.

A	"a"	"b"
1	2/3	1/3
2	1/4	3/4
3	3/5	2/5

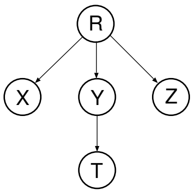
B	1	2	3
"a"	1/3	1/3	1/3
"b"	1/4	2/4	1/4

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) ¿Cuál es el mejor valor de X sabiendo que $R=1$?

a) Representación gráfica de la red:



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

$$\begin{aligned}
P(X,Y,Z \mid R) &= \frac{P(R,X,Y,Z)}{P(R)} \\
&= \frac{\sum_t P(R,T=t,X,Y,Z)}{P(R)} \\
&= \frac{\sum_t \cancel{P(R)}\ P(X \mid R)\ P(Y \mid R)\ P(Z \mid R)\ P(T=t \mid Y)}{\cancel{P(R)}} \\
&= P(X \mid R)\ P(Y \mid R)\ P(Z \mid R) \sum_t P(T=t \mid Y) \\
&= P(X \mid R)\ P(Y \mid R)\ P(Z \mid R)
\end{aligned}$$

c) El mejor valor de X sabiendo que $R=1$ se calcula:

Solución trivial: A partir de la tabla que define $P(X \mid R)$, $P(X=\texttt{"a"} \mid R=1) = 2/3$, $P(X=\texttt{"b"} \mid R=1) = 1/3$, por tanto el mejor valor de X es "a".

Solución indirecta: A partir del resultado del apartado b)

$$\begin{aligned}
P(X=\texttt{"a"} \mid R=1) &= \sum_{v,w \in \{\texttt{"a"},\texttt{"b"}\}} P(X=\texttt{"a"}, Y=v, Z=w \mid R=1) \\
&= \sum_{v,w \in \{\texttt{"a"},\texttt{"b"}\}} P(X=\texttt{"a"} \mid R=1)\ P(Y=v \mid R=1)\ P(Z=w \mid R=1) \\
&= P(X=\texttt{"a"} \mid R=1) \ = \ 2/3 \\
P(X=\texttt{"b"} \mid R=1) &= 1/3
\end{aligned}$$

El mejor valor de X sabiendo que $R=1$ es "a"