Aprendizaje automático y minería de datos
Examen final

6 de junio de 2017

Recuerda poner tu nombre y apellidos en todas las hojas
Buena Suerte

Problema 1 (1 punto). Se da el siguiente conjunto de datos:

x	petal length	sepal length	petal width	sepal width	Clase
$x^{(1)}$	6,0	2,2	5,0	1,5	virginica
$x^{(2)}$	5,0	2,3	3,3	1,0	versicolor
$x^{(3)}$	7,9	3,8	6,4	2,0	virginica
$x^{(4)}$	4,6	3,4	1,4	0,3	setosa
$x^{(5)}$	6,0	2,7	5,1	1,6	versicolor
$x^{(6)}$	5,0	3,2	1,2	0,2	setosa

Si antrenasemos un SVM con Kernel gaussiano $K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$, ¿cuál sería la fila $f^{(3)}$ de la matriz transformada para $\sigma = 3$?

Problema 2 (2 puntos). Partiendo de la información que tenemos sobre los cinco usuarios de una plataforma web: $u_1 = (1,2), u_2 = (1,3), u_3 = (3,1), u_4 = (5,4), u_5 = (6,5)$, describir los pasos seguidos en un procedimiento de clustering jerárquico ascendente, empleando la estrategia del amalgamamiento completo utilizando como distancia entre usarios la distancia Manhattan (también conocida como la norma L_1 o cityblock).

Problema 3 (2 puntos). Se da el conjunto de items $\mathcal{D} = \{abcdef, abcde, ab, ac, bc, ad\}$ sobre el alfabeto $\mathcal{I} = \{a, b, c, d, e, f\}$. Se pide argumentar si las reglas $d \to a$ y $abc \to de$ son representativas ($\tau = 0,2$ y $\gamma = 0,8$).

Problema 4 (2 puntos). Construir una red neuronal para la función booleana $f : \{0,1\}^3 \to \{0,1\}$ dada por la siguiente tabla:

x_1	x_2	x_3	f
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Problema 5 (1 punto). Una entidad financiera usa un sistema muy complejo (con decenas de atributos) para seleccionar unos pocos clientes que recibirían una oferta de préstamo de valor fijo (5000 euros). Como tienen contratado a un analista financiero, le encargan encontrar un predictor sencillo que sea capaz de filtrar los clientes en función de un único atributo: el sueldo anual. El analista optó por utilizar regresión logística. Para valorar la calidad del predictor, se usa un pequeño dataset con solicitudes de seis clientes: cuatro con crédito denegado y dos con crédito otorgado.

Nombre cliente	x	y	$h_\theta(x)$
Álvaro Menéndez Pelayo	30000	1	0.9
María Gómez Pérez	26000	0	0.7
Mateo Oriol Sierra	18000	0	0.3
Teresa Rodríguez Fernández	28000	0	0.8
Marina Bolado Suárez	14000	0	0.2
Raúl Alegría	22000	1	0.5

Se pide dibujar la matriz de confusión y la curva ROC (Receiver/Relative Operating Characteristics), y encontrar el AUC (*area under curve* = area bajo la curva). ¿Sería oportuno que la entidad financiera utilizara este predictor? Razonar la respuesta.

Problema 6 (2 puntos). Se da el lenguaje $L = \{w \in \{0,1\}^* \mid 010 \text{ is a subsequence of } w\}$. Se pide similar una ejecución del algoritmo L^* (Angluin) para inferir el automata finito determinista que acepta el lenguaje L.

Problema 7 (Pregunta bonus: 1 punto). ¿Cuál es el tema de la tesis doctoral de la profesora responsable de la asignatura? (la punctuación para esta pregunta dependerá del grado de precisión)

DEFINICIÓN 1. Dados $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$, elegimos $l^{(1)} = x^{(1)}, \ldots, l^{(m)} = x^{(m)}$.

Para un ejemplo $x^{(i)}$ en R^{p+1} construimos $f^{(i)}$ en R^{m+1}:

$f_0^{(i)} = 1$

$f_j^{(i)} = K(x^{(i)}, l^{(j)}) = K(x^{(i)}, x^{(j)}), \forall j \in \{1, \ldots, m\}$

Nota: en nuestro caso concreto, $p = 4$, $m = 6$.

DEFINICIÓN 2. En el caso del clustering jerárquico ascendente con amalgamiento completo, la distancia entre el cluster X y el cluster Y es la distancia máxima entre los puntos constituyentes:

$d(X, Y) = \text{máx}_{x \in X, y \in Y}\, d(x, y)$

DEFINICIÓN 3. Nota: la unión de conjuntos se denota por la justaposición de los mismos. Es decir, *abcd* es el conjunto $\{a, b, c, d\}$ y XY es la unión del conjunto X con el conjunto Y.

$F_\tau = \{X \subseteq \mathcal{I} \mid sup(X) \geq \tau\}$,
$FC_\tau = \{X \in F_\tau \mid \forall Z \supset X, sup(Z) < sup(X)\}$,
$FG_\tau = \{X \in F_\tau \mid \forall Y \subset X, sup(Y) > sup(X)\}$,
$RI_{\tau,\gamma} = \{X \in FC_\tau \mid \gamma * mxgs_{\tau,\gamma}(X) > mxs_\tau(X)\}$
$RR_{\tau,\gamma} = \{X \to Y \mid Y \neq \emptyset, X \cap Y = \emptyset, X \in FG_\tau, XY \in RI_{\tau,\gamma}, mxs_\tau(XY) < \gamma * sup(X) \leq sup(XY) < \gamma * mns_\tau(X)\}$
$mxs_\tau(X) = \text{máx}(\{sup(Z) \mid Z \in FC_\tau, Z \supset X\} \cup \{0\})$,
$mns_\tau(X) = \text{mín}(\{sup(Y) \mid Y \in FG_\tau, Y \subset X\} \cup \{\infty\})$,
$mxgs_{\tau,\gamma}(X) = \text{máx}(\{sup(Y) \mid Y \in FG_\tau, Y \subset X, sup(Y) \leq \frac{sup(X)}{\gamma}\} \cup \{0\})$.

DEFINICIÓN 4. La función de activación logística es

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

DEFINICIÓN 5. Para cada *threshold* k en [0,1] se calcula *True positives rate* (TPR) y *False positive rate* (FPR):

- TPR = $\frac{\text{casos buenos aceptados}}{\text{casos buenos}}$
- FPR = $\frac{\text{casos malos aceptados}}{\text{casos malos}}$

Luego se plotea TPR (en ordenadas) en función de FPR (en abscisas).