

Sistemas de Ayuda a la Decisión



1. Examen Teórico (3 ptos)

Nombre y Apellidos:.....

Índice

1. Enunciado	1
2. Anexo	4

1. Enunciado

1. Explica (0.5 ptos, 20 minutos):

- a) (0.1) Dada siguiente formalización de los conjuntos de entrenamiento, desarrollo y test para una tarea supervisada, rellena la parte

$$\begin{aligned}\mathcal{X}^{train} &= \{x^{\{1\}}, x^{\{2\}}, \dots x^{\{n\}}\} \\ \mathcal{X}^{dev} &= \{\underset{\{n+1\}}{X}, \underset{\{n+2\}}{X}, \dots, \underset{\{l\}}{X}\} \\ \mathcal{X}^{test} &= \{\underset{\{l+1\}}{X}, \underset{\{l+2\}}{X}, \dots, \underset{\{z\}}{X}\}\end{aligned}$$

n es tamaño de la muestra train, es decir, las instancias van de 1 a n, l-n+1 el tamaño del dev (es decir, las instancias que van de n+1 a l), y z-l-n+1 es el tamaño del test (es decir, las instancias que van de l a z)

$$\begin{aligned}\mathcal{Y}^{train} &= \{y^{\{1\}}, y^{\{2\}}, \dots y^{\{n\}}\} \\ \mathcal{Y}^{dev} &= \{\underset{\{n+1\}}{y}, \underset{\{n+2\}}{y}, \dots, \underset{\{l\}}{y}\} \\ \mathcal{Y}^{test} &= \{\underset{\{l+1\}}{y}, \underset{\{l+2\}}{y}, \dots, \underset{\{z\}}{y}\}\end{aligned}$$

Las Y contienen los valores reales

$$\begin{aligned}\mathcal{D}^{train} &= \{(x^{\{1\}}, y^{\{1\}}), (x^{\{2\}}, y^{\{2\}}), \dots (x^{\{n\}}, y^{\{n\}})\} \\ \mathcal{D}^{dev} &= \{(x^{\{n+1\}}, y^{\{n+1\}}), (x^{\{n+2\}}, y^{\{n+2\}}), \dots (x^{\{l\}}, y^{\{l\}})\} \\ \mathcal{D}^{test} &= \{(x^{\{l+1\}}, y^{\{l+1\}}), (x^{\{l+2\}}, y^{\{l+2\}}), \dots (x^{\{z\}}, y^{\{z\}})\}\end{aligned}$$

b) (0.1) ¿En que consiste la evaluación no honesta?

c) (0.1) ¿Por qué crees que cuando entrenamos varios modelos basados en distintos algoritmos donde para cada algoritmo tendremos que hacer un barrido de hiperparámetros es necesario dividir los datos en train (entrenamiento), development (desarrollo) y finalmente test? ¿Para qué se emplea cada uno?

d) (0.1) Identifica la mayor desventaja de un único Árbol de Decisión que se soluciona con el Random Forest y explica brevemente la razón por la que se soluciona con un Random Forest.

- e) (0.1) ¿Cúal(es) de los siguientes algoritmos funciona(n) **solo** con atributos numéricos? TACHA EL-LOS QUE SELECCIONES: Knn, Árboles de Decisión, Random Forest o Naive Bayes.

2. Ejercicio (0.5 ptos, 20 minutos): Selecciona de entre las fórmulas para calcular la precision la que creas más conveniente para estos datos y calcula su valor. Para ello crea y emplea la matriz de confusión:

	L-Sepalo	A-Sepalo	L-Petalo	A-Petalo	prediction	real
0	5.1	3.5	1.4	0.2	0	0
1	5	3.5	1.6	0.6	0	0
2	5.1	3.8	1.9	0.4	0	0
3	4.8	3	1.4	0.3	0	0
4	5.1	3.8	1.6	0.2	0	0
5	5	3.3	1.4	0.2	0	0
6	7	3.2	4.7	1.4	1	1
7	6.9	3.1	4.9	1.5	2	1
8	5.5	2.3	4	1.3	1	1
9	6.5	2.8	4.6	1.5	2	1
10	5.1	2.5	3	1.1	1	1
11	5.7	2.8	4.1	1.3	1	1
12	6.3	3.3	6	2.5	2	2
13	5.8	2.7	5.1	1.9	2	2
14	7.1	3	5.9	2.1	2	2

3. (0.25 ptos, 10 minutos) Suponiendo una tarea de clasificación de SPAMS, explica brevemente los preprocesos que aplicarías, como quedaría el mensaje tras esos preprocesos y calcula el BOW del SMS con identificador Id 1 en su versión one-hot-vector (es decir, solo ceros y unos) de datos de SMS-s que encontrarás en el Anexo (nota: el diccionario/Vector también está en el anexo).
4. (1 pto, 30 minutos) Empleando para el entrenamiento los datos que encontrarás en el Anexo sobre el juego del Tennis. Predice para la siguiente instancia su clase. Emplea el algoritmo que deseas y justifíca tu elección. A la hora de tomar tu decisión, ten en cuenta que los datos son categoriales, que la muestra es muy pequeña y al ser tan pequeña existe un serio riesgo de sobreajuste (overfitting) y que luego no generalice correctamente:

Outlook	Temperature	Humidity	PlayTennis
Sunny	cold	High	No

5. (0.75 ptos, 20 minutos) Empleando para el entrenamiento los datos que encontrarás en el Anexo sobre el juego del Tennis. ¿Cuál de los nodos es mejor candidato para nodo raíz en un árbol de decisión, la temperatura o la humedad? (Emplea la Ganancia de Información para justificar tu respuesta).

2. Anexo

Id	SMS	Clase Real	Clase Pred.
1	Urgent! Finished class where are you.	0	1
2	Congratulations! One year older! I am calling U...	0	1
3	Sorry, I'll call later	0	0
4	Remember U owe me 20 pounds	0	1
5	K. Did you call me just now ah?	0	0
6	Ok i am on the way to home. Do you offer me dinner tonight?	0	1

Cuadro 1: Ejemplos de un Conjunto de SMS

0	1	2	3	4	5	6	7
!	,	.	?	congratulation	2	20	2000
8	9	10	11	12	13	14	15
3510i	4	500	accomodate	award	black	busy	call
16	17	18	19	20	21	22	23
cash-balance	chat	class	congratulation	contact	current	dear	dinner
24	25	26	27	28	29	30	31
draw	dream	finish	friday	friend	go	good	guess
32	33	34	35	36	37	38	39
he	hear	home	i	iphone	late	line	logo
40	41	42	43	44	45	46	47
lover	me	motorola	msg	new	nokia	not	now
48	49	50	51	52	53	54	55
number	offer	ok	old	one	owe	phone	pls
56	57	58	59	60	61	62	63
pound	private	prize	remember	rude	send	show	sorry
64	65	66	67	68	69	70	71
thank	today	tomorrow	tonight	try	u	ur	urgent
72	73	74	75	76	77	78	79
video	want	way	we	where	while	win	worry
80	81	82					
xmas	year	you					

Cuadro 2: Vector BOW para llenar con los datos de los SMS

Outlook	Temperature	Humidity	PlayTennis
Sunny	hot	High	No
Sunny	hot	High	No
Overcast	hot	High	Yes
Rainy	mild	High	Yes
Rainy	cold	Normal	Yes
Rainy	cold	Normal	No
Overcast	cold	Normal	Yes
Sunny	mild	High	No
Sunny	cold	Normal	Yes
Rainy	mild	Normal	Yes
Sunny	mild	Normal	Yes
Overcast	mild	High	Yes
Overcast	hot	Normal	Yes
Rainy	mild	High	No

Cuadro 3: Datos del tenis

- a) (0.1) Dada siguiente formalización de los conjuntos de entrenamiento, desarrollo y test para una tarea supervisada, rellena la parte

$$\mathcal{X}^{train} = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$$

$$\mathcal{X}^{dev} = \{\overset{\{n+1\}}{x}, \overset{\{n+2\}}{x}, \dots, \overset{\{l\}}{x}\}$$

$$\mathcal{X}^{test} = \{\overset{\{l+1\}}{x}, \overset{\{l+2\}}{x}, \dots, \overset{\{z\}}{x}\}$$

n es tamaño de la muestra train, es decir, las instancias van de 1 a n, l-n+1 el tamaño del dev (es decir, las instancias que van de n+1 a l), y z-l-n+1 es el tamaño del test (es decir, las instancias que van de l a z)

$$Y^{train} = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$$

$$Y^{dev} = \{\overset{\{n+1\}}{y}, \overset{\{n+2\}}{y}, \dots, \overset{\{l\}}{y}\}$$

$$Y^{test} = \{\overset{\{l+1\}}{y}, \overset{\{l+2\}}{y}, \dots, \overset{\{z\}}{y}\}$$

Las Y contienen los valores reales

$$\mathcal{D}^{train} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$$

$$\mathcal{D}^{dev} = \{(x^{(n+1)}, y^{(n+1)}), (x^{(n+2)}, y^{(n+2)}), \dots, (x^{(l)}, y^{(l)})\}$$

$$\mathcal{D}^{test} = \{(x^{(l+1)}, y^{(l+1)}), (x^{(l+2)}, y^{(l+2)}), \dots, (x^{(z)}, y^{(z)})\}$$

- b) (0.1) ¿En qué consiste la evaluación no honesta?

Entrenar y evaluar con el mismo conjunto de datos

- c) (0.1) ¿Por qué crees que cuando entrenamos varios modelos basados en distintos algoritmos donde para cada algoritmo tendremos que hacer un barrido de hiperparámetros es necesario dividir los datos en train (entrenamiento), development (desarrollo) y finalmente test? ¿Para qué se emplea cada uno?

Para conseguir que no haya overfitting y que el modelo sea capaz de generalizar. El conjunto de entrenamiento se emplea para entrenar el modelo, el conjunto de desarrollo se emplea para ajustar los hiperparámetros y el conjunto de test se emplea para evaluar el modelo.

- d) (0.1) Identifica la mayor desventaja de un único Árbol de Decisión que se soluciona con el Random Forest y explica brevemente la razón por la que se soluciona con un Random Forest.

El overfitting. Se soluciona con el Random Forest porque se entrena con distintos subconjuntos de datos y distintas variables.

- e) (0.1) ¿Cuál(es) de los siguientes algoritmos funciona(n) **solo** con atributos numéricos? TACHA EL-LOS QUE SELECCIONES: Knn, Árboles de Decisión, Random Forest o Naive Bayes.

KNN y Naive Bayes

2. Ejercicio (0.5 ptos, 20 minutos): Selecciona de entre las fórmulas para calcular la precisión la que creas más conveniente para estos datos y calcula su valor. Para ello crea y emplea la matriz de confusión:

	L-Sepalo	A-Sepalo	L-Petalo	A-Petalo	prediction	real
0	5.1	3.5	1.4	0.2	0	0
1	5	3.5	1.6	0.6	0	0
2	5.1	3.8	1.9	0.4	0	0
3	4.8	3	1.4	0.3	0	0
4	5.1	3.8	1.6	0.2	0	0
5	5	3.3	1.4	0.2	0	0
6	7	3.2	4.7	1.4	1	1
7	6.9	3.1	4.9	1.5	2	1
8	5.5	2.3	4	1.3	1	1
9	6.5	2.8	4.6	1.5	2	1
10	5.1	2.5	3	1.1	1	1
11	5.7	2.8	4.1	1.3	1	1
12	6.3	3.3	6	2.5	2	2
13	5.8	2.7	5.1	1.9	2	2
14	7.1	3	5.9	2.1	2	2

Pred	
Real	Pred
0	0 ~
0	6 0
~	0 7
Real	Pred
1	1 ~
1	4 2
~	0 9
Real	Pred
2	2 ~
2	3 0
~	2 10

TP	TN	FP	FN
0 6	7	0	0
1 4	9	0	2
2 3	10	2	0

3. (0.25 ptos, 10 minutos) Suponiendo una tarea de clasificación de SPAMS, explica brevemente los preprocesos que aplicarías, como quedaría el mensaje tras esos preprocesos y calcula el BOW del SMS con identificador Id 1 en su versión one-hot-vector (es decir, solo ceros y unos) de datos de SMS-s que encontrarás en el Anexo (nota: el diccionario/Vector también está en el anexo).

Id	SMS	Clase Real	Clase Pred.
1	Urgent! Finished class where are you.	0	1
2	Congratulations! One year older! I am calling U...	0	1
3	Sorry, I'll call later	0	0
4	Remember U owe me 20 pounds	0	1
5	K. Did you call me just now ah?	0	0
6	Ok i am on the way to home. Do you offer me dinner tonight?	0	1

1 - Tokenizar

2 - Normalizar

3 - Stop Words

4 - Lematizar

0	1	2	3	4	5	6	7
!	,	.	?	congratulation	2	20	2000
1	0	1	0	0	0	0	0
8	9	10	11	12	13	14	15
3510i	4	500	accomodate	award	black	busy	call
0	0	0	0	0	0	0	0
16	17	18	19	20	21	22	23
cash-balance	chat	class	congratulation	contact	current	dear	dinner
0	0	1	0	0	0	0	0
24	25	26	27	28	29	30	31
draw	dream	finish	friday	friend	go	good	guess
0	0	1	0	0	0	0	0
32	33	34	35	36	37	38	39
he	hear	home	i	iphone	late	line	logo
0	0	0	0	0	0	0	0
40	41	42	43	44	45	46	47
lover	me	motorola	msg	new	nokia	not	now
0	0	0	0	0	0	0	0
48	49	50	51	52	53	54	55
number	offer	ok	old	one	owe	phone	pls
0	0	0	0	0	0	0	0
56	57	58	59	60	61	62	63
pound	private	prize	remember	rude	send	show	sorry
0	0	0	0	0	0	0	0
64	65	66	67	68	69	70	71
thank	today	tomorrow	tonight	try	u	ur	urgent
0	0	0	0	0	0	0	1
72	73	74	75	76	77	78	79
video	want	way	we	where	while	win	worry
0	0	0	0	1	0	0	0
80	81	82					
xmas	year	you					
0	0	1					

4. (1 pto, 30 minutos) Empleando para el entrenamiento los datos que encontrarás en el Anexo sobre el juego del Tennis. Predice para la siguiente instancia su clase. Emplea el algoritmo que deseas y justifica tu elección. A la hora de tomar tu decisión, ten en cuenta que los datos son categóricos, que la muestra es muy pequeña y al ser tan pequeña existe un serio riesgo de sobreajuste (overfitting) y que luego no generalice correctamente:

Outlook	Temperature	Humidity	PlayTennis
Sunny	cold	High	No

	Outlook	Temperature	Humidity	PlayTennis
1	Sunny	hot	High	No
2	Sunny	hot	High	No
3	Overcast	hot	High	Yes
4	Rainy	mild	High	Yes
5	Rainy	cold	Normal	Yes
6	Rainy	cold	Normal	No
7	Overcast	cold	Normal	Yes
8	Sunny	mild	High	No
9	Sunny	cold	Normal	Yes
10	Rainy	mild	Normal	Yes
11	Sunny	mild	Normal	Yes
12	Overcast	mild	High	Yes
13	Overcast	hot	Normal	Yes
14	Rainy	mild	High	No

Cuadro 3: Datos del tenis

	Sunny	Overcast	Rainy	hot	mild	cold	High	Wind
Play 9	2	4	3	2	4	3	3	6
Play 5	3	0	2	2	2	1	4	1
	14			14			14	

$$P(\text{Play} | \text{Sunny, cold, high})$$

$$P(\text{Sunny} | \text{Play}) \cdot P(\text{cold} | \text{Play}) \cdot P(\text{High} | \text{Play}) \cdot P(\text{Play})$$

$$\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$$

$$\underline{0.0258}$$

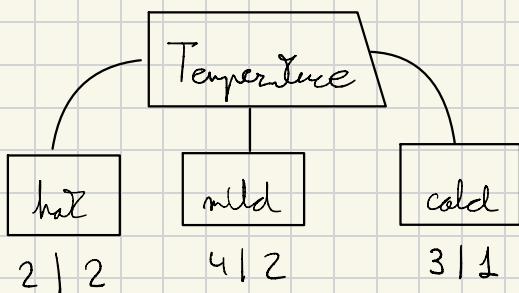
$$P(\text{!Play} | \text{Sunny, cold, high})$$

$$P(\text{Sunny} | \text{!Play}) \cdot P(\text{cold} | \text{!Play}) \cdot P(\text{High} | \text{!Play}) \cdot P(\text{Play})$$

$$\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{5}{14}$$

$$\underline{0.0342}$$

5. (0.75 ptos, 20 minutos) Empleando para el entrenamiento los datos que encontrarás en el Anexo sobre el juego del Tennis. ¿Cuál de los nodos es mejor candidato para nodo raíz en un árbol de decisión, la temperatura o la humedad? (Emplea la Ganancia de Información para justificar tu respuesta).



Outlook	Temperature	Humidity	PlayTennis
Sunny	hot	High	No
Sunny	hot	High	No
Overcast	hot	High	Yes
Rainy	mild	High	Yes
Rainy	cold	Normal	Yes
Rainy	cold	Normal	No
Overcast	cold	Normal	Yes
Sunny	mild	High	No
Sunny	cold	Normal	Yes
Rainy	mild	Normal	Yes
Sunny	mild	Normal	Yes
Overcast	mild	High	Yes
Overcast	hot	Normal	Yes
Rainy	mild	High	No

Cuadro 3: Datos del tenis

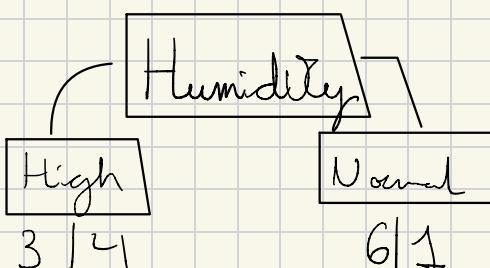
$$EH = \left[-\frac{2}{4} \cdot \log_3\left(\frac{2}{4}\right) - \frac{2}{4} \cdot \log_3\left(\frac{2}{4}\right) \right] + \left[-\frac{4}{6} \cdot \log_3\left(\frac{4}{6}\right) - \frac{2}{6} \cdot \log_3\left(\frac{2}{6}\right) \right] + \left[-\frac{3}{4} \cdot \log_3\left(\frac{3}{4}\right) - \frac{1}{4} \cdot \log_3\left(\frac{1}{4}\right) \right]$$

$$(0,246 + 0,33) \quad (0,196 + 0,315)$$

$$0,63 \quad 0,576 \quad 0,551$$

$$EH = \frac{4}{14} \cdot 0,63 + \frac{6}{14} \cdot 0,576 + \frac{4}{14} \cdot 0,551 = 0,573$$

$$IG = 1 - 0,679 = \underline{0,427}$$



$$EH = \left[-\frac{3}{7} \log_2\left(\frac{3}{7}\right) - \frac{4}{7} \log_2\left(\frac{4}{7}\right) \right] + \left[-\frac{6}{7} \log_2\left(\frac{6}{7}\right) - \frac{1}{7} \log_2\left(\frac{1}{7}\right) \right]$$

$$(0,52 + 0,46) + (0,19 + 0,4)$$

$$0,98 \quad 0,59$$

$$EH = \frac{7}{14} \cdot 0,98 + \frac{7}{14} \cdot 0,59 = 0,785$$

$$IG = 1 - 0,785 = \underline{0,215}$$