

Sistemas de Ayuda a la Decisión



1. Examen Teórico (2 ptos)

Nombre y Apellidos:

Índice

1. Enunciado	1
2. Anexo	4

1. Enunciado

1. Explica (0.75 ptos, 40 minutos):

- a) Dado el siguiente pseudocódigo del KNN (0.15 ptos):

Algorithm 1 Predicción-Knn($k, d_{card}, \mathcal{D}^{train}, \hat{x}$)

```
Aux = []
for each  $x^{\{t\}} \in \mathcal{D}^{train}$  do
    Aux = Aux  $\oplus$   $< d(x^{\{t\}}, \hat{x}), c >$ 
end for
ordenar(Aux)
C_Clases = []
for  $k = 1$  to  $K$  do
    actualizar(C_Clases,  $c^{\{k\}}$ )
end for
clasePred = max(C_Clases)
```

Describe brevemente cada variable y cada paso del algoritmo.

Ejemplo ¿Qué representa c ? c representa la clase de

¿Cuál es el rol de Aux en el algoritmo? Aux almacena

¿Qué representa $x^{\{t\}}$? $x^{\{t\}}$ representa

¿Qué representa el hiperparámetro K ? K representa

¿Qué almacena C_Clases ? C_Clases almacena.....

Algoritmo en pseudocódigo castellano:

POR CADA elemento de

Almacenar

POR CADA k de 1 a K

- b) Indica cómo modificarías el código para que el voto de los K elementos no valiese igual (0.15 pto)

- c) Explica lo que es el overfitting. ¿Por qué los árboles de decisión sufren más de él y cómo lo solucionan los Random Forest? (0.15 ptos).

- Overfitting es cuando el modelo se ajusta demasiado lo que hace que no generalice correctamente
- Un árbol de decisión puede crecer mucha en profundidad que es lo que genera overfitting
- El random forest crea varios árboles y combinan sus predicciones para mitigar el overfitting

- d) A través del ejemplo del bibliotecario y el granjero se ejemplifica claramente la importancia de la probabilidad a priori en la formulación de Naïve Bayes. X es **tímido, retraído y dispuesto a ayudar**. X muestra **poco interés por la gente y por la realidad que le rodea**. X es **pacífico, ordenado y siente pasión por los detalles**. X necesita un **ambiente ordenado y estructurado**. ¿Cuál es la respuesta que suele dar la gente a la pregunta ”¿Qué es X Bibliotecario o granjero?”. Explica brevemente por qué no es correcta la respuesta habitual y qué papel juega la probabilidad a priori en la fórmula de Naive Bayes (0.15 ptos).

- e) Explica la fórmula de Naive Bayes (0.15 ptos)

$$Prob(H|E) = \frac{Prob(H)Prob(E|H)}{Prob(E)}$$

empleando el ejemplo del Bibliotecario y el granjero.

2. (0.25 ptos, 15 minutos) Suponiendo una tarea de clasificación de SPAMS, explica brevemente los preprocesos que aplicarías, como quedaría el mensaje tras esos preprocesos y calcula el tf-idf del SMS con identificador Id 1 de datos de SMS-s que encontrarás en el Anexo (nota: el diccionario/Vector también está en el anexo).
3. (1 pto, 45 minutos) Digamos que disponemos de un documento que habla de turismo y en el que se mencionan atracciones turísticas de una ciudad, específicamente el **Ayuntamiento** y el **Prado** y queremos saber a qué ciudad se refiere, sabiendo que las ciudades entre las que queremos clasificar están **Toledo, Segovia, Bilbao, Donostia y Madrid**. Para poder realizar dicha predicción disponemos de información de otros documentos y éstos se han anotado con la ciudad a la que se refieren (Anexo, Cuadro 3 instancias de documentos, las menciones de atracciones contenidas en cada uno y la ciudad a la que se refiere el documento). Se pide predecir a qué ciudad se está refiriendo el documento, aplicando Naïve Bayes (0.15 ptos el cálculo de las probabilidades a priori, 0.85 ptos el cálculo de las probabilidades de cada hipótesis y determinación de la clase más probable para la nueva instancia [**Ayuntamiento,Prado**]). Repetir el ejercicio con el Cuadro3 y siendo la instancia [**Pintxos,Museo Etnográfico**]

2. Anexo

Id	SMS	Clase Real	Clase Pred.
1	Sorry!, I will call later	0	0
2	Urgent! Finished class where are you.	0	1
3	Congratulations! One year older! I am calling U...	0	1
4	Remember U owe me 20 pounds	0	1
5	Sorry, 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 tf-idf para llenar con los datos del primer SMS

doc1	Alcazar	Catedral	Greco				Toledo
doc2	Alcazar	Catedral					Toledo
doc3	Alcazar	Acueducto	Catedral				Segovia
doc4	Acueducto						Segovia
doc5	Acueducto	Moneda	Cochinillo				Segovia
doc6	Pintxos	Guggenheim	Ayuntamiento	Pintxos	Pintxos	Pintxos	Bilbao
doc7	Prado	Sofia	Thyssen				Madrid
doc8	Prado	Cervantes					Madrid
doc9	Sofia	Thyssen					Madrid
doc10	Palacio	Retiro	Botánico				Madrid
doc11	Acuario	Pintxos	Pintxos	Pintxos	Pintxos		Donostia

Cuadro 3: Datos sobre documentos sobre atracciones turísticas

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Guggenheim
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
Acuario							

Cuadro 4: Diccionario de 17 Atracciones (atributos para Cuadro3)

2. (0.25 ptos, 15 minutos) Suponiendo una tarea de clasificación de SPAMS, explica brevemente los preprocessos que aplicarías, como quedaría el mensaje tras esos preprocessos y calcula el tf-idf del SMS con identificador Id 1 de datos de SMS-s que encontrarás en el Anexo (nota: el diccionario/Vector también está en el anexo).

1. Tokenizar → Esperar signos y palabras
2. Normalizar → Quitar mayúsculas, acentos, simbolos raros.
3. Stop Words → Eliminar palabras sin valor
4. Lematizar → Quedarse con la raíz de las palabras

Id	SMS	Clase Real	Clase Pred.
1	Sorry!, I will call later	0	0
2	Urgent! Finished class where are you.	0	1
3	Congratulations! One year older! I am calling U...	0	1
4	Remember U owe me 20 pounds	0	1
5	Sorry, 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

0	1	2	3	4	5	6	7
!	,	.	?	congratulation	2	20	2000
0,0 5	0,0 8	0	0	0	0	0	0
8	9	10	11	12	13	14	15
3510i	4	500	accommodate	award	black	busy	call
0	0	0	0	0	0	0	0,43
16	17	18	19	20	21	22	23
cash-balance	chat	class	congratulation	contact	current	dear	dinner
0	0	0	0	0	0	0	0
24	25	26	27	28	29	30	31
draw	dream	finish	friday	friend	go	good	guess
0	0	0	0	0	0	0	0
32	33	34	35	36	37	38	39
he	hear	home	i	iphone	late	line	logo
0	0	0	0,05	0	0,43	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	0,07
64	65	66	67	68	69	70	71
thank	today	tomorrow	tonight	try	u	ur	urgent
0	0	0	0	0	0	0	0
72	73	74	75	76	77	78	79
video	want	way	we	where	while	win	worry
0	0	0	0	0	0	0	0
80	81	82					
xmas	year	you					
0	0	0					

1 2 3 4 5 6
→ sorry ! , i call late

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{2} \end{aligned} \quad \left. \right\} 0,08$$

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{3} \end{aligned} \quad \left. \right\} 0,05$$

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{2} \end{aligned} \quad \left. \right\} 0,08$$

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{3} \end{aligned} \quad \left. \right\} 0,05$$

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{3} \end{aligned} \quad \left. \right\} 0,05$$

$$\begin{aligned} \text{tf} &= \frac{1}{6} \\ \text{idf} &= \log \frac{6}{1} \end{aligned} \quad \left. \right\} 0,13$$

3. (1 pto, 45 minutos) Digamos que disponemos de un documento que habla de turismo y en el que se mencionan atracciones turísticas de una ciudad, específicamente el **Ayuntamiento** y el **Prado** y queremos saber a qué ciudad se refiere, sabiendo que las ciudades entre las que queremos clasificar están **Toledo**, **Segovia**, **Bilbao**, **Donostia** y **Madrid**. Para poder realizar dicha predicción disponemos de información de otros documentos y éstos se han anotado con la ciudad a la que se refieren (Anexo, Cuadro 3 instancias de documentos, las menciones de atracciones contenidas en cada uno y la ciudad a la que se refiere el documento). Se pide predecir a qué ciudad se está refiriendo el documento, aplicando Naïve Bayes (0.15 ptos el cálculo de las probabilidades a priori, 0.85 ptos el cálculo de las probabilidades de cada hipótesis y determinación de la clase más probable para la nueva instancia [**Ayuntamiento, Prado**]). Repetir el ejercicio con el Cuadro3 y siendo la instancia [**Pintxos, Museo Etnográfico**]

doc1	Alcazar	Catedral	Greco				Toledo	} 2/22
doc2	Alcazar	Catedral					Toledo	
doc3	Alcazar	Acueducto	Catedral				Segovia	} 3/11
doc4	Acueducto						Segovia	
doc5	Acueducto	Moneda	Cochinillo				Segovia	} 3/11
doc6	Pintxos	Gugenheim	Ayuntamiento	Pintxos	Pintxos	Pintxos	Bilbao	
doc7	Prado	Sofia	Thyssen				Madrid	} 4/11
doc8	Prado	Cervantes					Madrid	
doc9	Sofia	Thyssen					Madrid	} 4/11
doc10	Palacio	Retiro	Botánico				Madrid	
doc11	Acuario	Pintxos	Pintxos	Pintxos	Pintxos		Donostia	1/11

Toledo

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Gugenheim
2	3	3	1	1	1	1	1
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
1	4	4	4	4	1	1	1
Acuario							
1							

22

Segovia

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Gugenheim
1	2	2	4	2	2	1	1
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
1	4	6	1	8	1	1	1
Acuario							
1							

24

Bilbao

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Gugenheim
1	1	1	1	1	1	5	2
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
1	1	1	1	1	1	1	2
Acuario							
1							

23

Madrid

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Gugenheim
1	1	1	1	1	1	1	1
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
3	2	3	3	2	2	2	1
Acuario							
1							

28

Donostia

Greco	Alcazar	Catedral	Acueducto	Moneda	Cochinillo	Pintxos	Gugenheim
1	1	1	1	1	1	5	1
Prado	Cervantes	Sofia	Thyssen	Retiro	Botánico	Palacio	Ayuntamiento
1	1	1	1	1	1	1	1
Acuario							
2							

22

3. (1 pto, 45 minutos) Digamos que disponemos de un documento que habla de turismo y en el que se mencionan atracciones turísticas de una ciudad, específicamente el **Ayuntamiento** y el **Prado** y queremos saber a qué ciudad se refiere, sabiendo que las ciudades entre las que queremos clasificar están **Toledo**, **Segovia**, **Bilbao**, **Donostia** y **Madrid**. Para poder realizar dicha predicción disponemos de información de otros documentos y éstos se han anotado con la ciudad a la que se refieren (Anexo, Cuadro 3 instancias de documentos, las menciones de atracciones contenidas en cada uno y la ciudad a la que se refiere el documento). Se pide predecir a qué ciudad se está refiriendo el documento, aplicando Naïve Bayes (0.15 ptos el cálculo de las probabilidades a priori, 0.85 ptos el cálculo de las probabilidades de cada hipótesis y determinación de la clase más probable para la nueva instancia [**Ayuntamiento, Prado**]). Repetir el ejercicio con el Cuadro3 y siendo la instancia [**Pintxos, Museo Etnográfico**]

$P(\text{Toledo} | \text{Ayuntamiento, Prado})$

$$P(\text{Ayuntamiento} | \text{Toledo}) \cdot P(\text{Prado} | \text{Toledo}) \cdot P(\text{Toledo})$$

$$\frac{1}{22} \cdot \frac{1}{22} \cdot \frac{2}{11} = 0,000375$$

$P(\text{Segovia} | \text{Ayuntamiento, Prado})$

$$P(\text{Ayuntamiento} | \text{Segovia}) \cdot P(\text{Prado} | \text{Segovia}) \cdot P(\text{Segovia})$$

$$\frac{1}{24} \cdot \frac{1}{24} \cdot \frac{3}{11} = 0,000473$$

$P(\text{Bilbao} | \text{Ayuntamiento, Prado})$

$$P(\text{Ayuntamiento} | \text{Bilbao}) \cdot P(\text{Prado} | \text{Ayuntamiento}) \cdot P(\text{Bilbao})$$

$$\frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{11} = 0,000344$$

$P(\text{Madrid} | \text{Ayuntamiento, Prado})$

$$P(\text{Ayuntamiento} | \text{Madrid}) \cdot P(\text{Prado} | \text{Madrid}) \cdot P(\text{Madrid})$$

$$\frac{1}{28} \cdot \frac{3}{28} \cdot \frac{4}{11} = 0,00139$$

$P(\text{Donostia} | \text{Ayuntamiento, Prado})$

$$P(\text{Ayuntamiento} | \text{Donostia}) \cdot P(\text{Prado} | \text{Donostia}) \cdot P(\text{Donostia})$$

$$\frac{1}{22} \cdot \frac{1}{22} \cdot \frac{1}{11} = 0,000188$$