

Grado en Ingeniería Informática

Explotación de la Información Módulo 4. Clasificación y Agrupamiento de Información

Antonio Ferrández Rodríguez



UNIVERSIDAD DE ALICANTE



Grupo de Procesamiento
del Lenguaje y Sistemas
de Información

Índice

1. Sistemas de clasificación de información
2. Clasificación basada en vocabulario
3. Clasificación utilizando árboles de decisión
4. Clasificación utilizando sistemas de reglas
5. Problema del *overfitting*. Sistemas de poda
6. Part of speech tagging
7. Sistemas de agrupamiento de información
8. Sistemas de agrupamiento de información en la Recuperación de Información
9. Sistemas de agrupamiento de información particionales (algoritmo *k-mean*)
10. Sistemas de agrupamiento de información jerárquicos

1. Sistemas de clasificación de Información

Clasificación automática/*automated classification*:

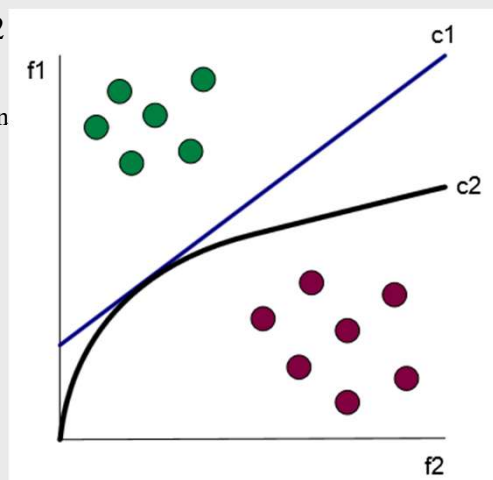
- Asignación de una categoría predefinida disjunta
 - Distinto del proceso de categorización (*categorization*):
 - # Se permite la asignación de más de una clase, etiqueta o categoría para cada instancia: p.ej. asignar temáticas a libros
- Resultado: ontologías, taxonomías, jerarquías, vocabularios controlados o tesauros
- Problemas: precisión, consistencia, etc.
- Aplicaciones:
 - RI como un problema de clasificación con las categorías documento relevante/no relevante
 - Detección de spam o detección de páginas con contenido violento
 - Detección de autor (*Authorship attribution*)
 - Part of speech tagging
 - Fluency ranking en generación de texto

3

1. Sistemas de clasificación de Información

■ Objetivo:

- Características f_1 y f_2
- Dos clases:
 - # Maximizar separación entre las clases



4

1. Sistemas de clasificación de Información

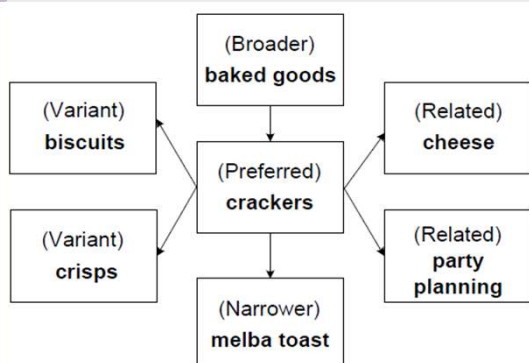
Técnicas:

- Técnicas basadas en vocabulario:
 - Usan un tesoro o diccionario para determinar aquellos términos y sus variantes asociados a cada categoría
 - Problema: ambigüedad del lenguaje
- Árboles de decisión
- Basadas en reglas
- Estadísticas: co-ocurrencia de términos, redes neuronales, etc.

5

2. Sist. clasificación información basadas en vocabulario

- Ej.: categoría comida



Example

Personal Digital Assistant

Synonyms

Handheld Computer

"Alternate" Spellings

Persenal Digitel Asistent

Abbreviations / Acronyms

PDA

Broader Terms

Wireless, Computers

Narrower Terms

PalmPilot, PocketPC

Related Terms

WindowsCE, Cell Phones

6

2. Sist. clasificación información basadas en vocabulario

Ejercicio 1:

- Obtener las reglas y vocabulario para las siguientes categorías de tipo de pregunta de los sistemas de búsqueda de respuesta. Utilizad a modo de ejemplo las preguntas que aparecen en la siguiente transparencia:
- En la siguiente URL se pueden encontrar ayudas de sinónimos y relaciones semánticas:

<http://adimen.si.chu.es/cgi-bin/wei/public/wei.consult.perl>

7

2. Sist. clasificación información basadas en vocabulario

entidad persona	¿como se llama el hijo de kim il sung?
entidad persona	¿quien es el creador de "doctor snuggles"?
entidad persona	¿quien es el lider bosnio?
entidad persona	¿quien fue la ganadora del torneo de wimbledon?
entidad persona	¿que presidente de corea del norte murio a los 82 años de edad?
entidad persona	¿quien es el presidente del parlamento europeo?
entidad persona	¿quien es el líder del sinn fein?
entidad persona	¿quien es el mayor exportador europeo de aceite de oliva?
entidad persona	¿quien escribio "star trek"?
entidad persona	¿quien es el presidente de la republica de italia?
entidad persona	¿quien ostenta el poder en pyongyang?
entidad persona	¿quien dirigio "con la muerte en los talones"?
entidad persona	¿quien es el presidente de rusia?
entidad persona	¿quien es el presidente italiano de asuntos exteriores?
entidad persona	¿quien es el entrenador del equipo nacional de futbol noruego?
entidad persona	¿quien es el director de la cia?
entidad persona	¿como se llamaba el cantante y lider de nirvana?
entidad persona	¿quien es el presidente de la republica francesa?
entidad persona	¿que primer ministro frances se suicido en los años 90?
entidad persona	¿quien es el presidente de peru?
entidad persona	¿que presidente ruso asistio a la reunion del g7 en napoles?
entidad persona	¿a que primer ministro abrio la fiscalia de milan un sumario por corrupcion?
entidad persona	¿quien proyecto la construccion de la catedral de san pedro?
entidad persona	¿como se llama el jefe de gobierno de australia?
entidad persona	¿como se llama el sucesor del gatt?
entidad persona	¿quien es el presidente de yugoslavia?
entidad persona	¿que ciudadano britanico recibio 50 latigazos en qatar?

entidad abreviat ¿cuales son las siglas del fondo mundial para la proteccion de la naturaleza?
entidad abreviat ¿cual es el acronimo de amnistia internacional?

8

Explotación de la información, Clasificación y Agrupamiento de Información

2. Sist. clasificación información basadas en vocabulario

entidad objeto	¿cual es la anterior moneda argentina?
entidad objeto	¿de que obtendra microsoft la licencia de sun?
entidad objeto	¿cual es el nombre de estandar europeo de comunicaciones moviles digitales?
entidad objeto	¿que produce la compania victorinox?
entidad objeto	¿que produce mico?
entidad objeto	¿cual es el simbolo de paris?
entidad objeto	¿que tecnologia produce leica?
entidad objeto	¿como se llama el ferry naufragado en suecia en 1994?
entidad objeto	¿contra que choco el titanic?
entidad objeto	¿cual es el simbolo de liderazgo del giro de italia?
entidad objeto	¿que fue levantado el 13 de agosto de 1961?
entidad objeto	¿cual es la moneda iraki?
entidad objeto	nombre un edificio envuelto por christo.
entidad objeto	nombre una pelicula en la que se hayan usado animaciones por ordenador.
entidad objeto	¿que deporte practica adrian mutu?
entidad objeto	¿que alfabeto tiene solo cuatro letras "a, c, g, y t"?
entidad objeto	¿que plataforma estaba acampada en el paseo de la castellana de madrid?
entidad objeto	¿a que enfermedad corresponden las siglas rsi?
entidad objeto	¿que tipo de dolencia es caracteristica del rsi?
entidad objeto	¿que vitaminas ayudan en la lucha contra el cancer?
entidad objeto	¿que fruta tiene vitamina c?
entidad objeto	¿con el nombre de que enfermedad se corresponde el acronimo bse?
entidad objeto	¿que submarino choco con un buque en el canal de la mancha el 16 de febrero de 1995?
entidad objeto	¿en que epoca del año desaparecio jurgen schneider al producirse la bancarrota de su empres?
entidad objeto	¿que premio gano pulp fiction en el festival de cine de cannes?
entidad objeto	¿que nuevo canal de television gay aparecio en francia el 25 de octubre de 2004?
entidad objeto	¿cual es la ultima letra del alfabeto fonetico de la otan?
entidad objeto	¿con que pelicula marlee matlin gano un oscar?
entidad objeto	¿que huracan azoto la isla de cozumel?
entidad objeto	nombre una pelicula en la que haya participado kirk douglas en el periodo de 1946 a 1960.
entidad objeto	de el nombre de alguien que haya ganado el premio nobel de literatura entre 1945 y 1990.
entidad objeto	¿con que planeta choco el cometa shoemaker-levy?
entidad objeto	¿cual es la palabra alemana mas larga?
entidad objeto	¿como se llama la moneda de letonia?
entidad objeto	¿en que calle vive el primer ministro britanico?

9

Explotación de la información, Clasificación y Agrupamiento de Información

2. Sist. clasificación información basadas en vocabulario

numerico econo	¿cuanto valen 10 pesos?
numerico econo	¿quanto costo el tunel del canal?
numerico econo	¿que gasto se ha programado en virtud del ifop en el periodo 1994-1999 para la renovacion de
numerico econo	¿a cuanto ascendieron los beneficios del grupo fines de electronica y comunicaciones nokia e
numerico econo	¿cuanto reclama el sevilla fc a diego maradona?
numerico econo	¿a cuanto asciende el premio para la ganadora de wimbledon?
numerico econo	¿cual era el valor aproximado de la carga de un galeon del siglo xvi?
numerico econo	¿quanto dinero gana anualmente el narcotrafico?
numerico econo	¿cual es el presupuesto de la interpol?
numerico econo	¿a cuanto asciende la multa que se le impuso a italia por superar la cuota de produccion de le
numerico econo	¿a cuanto ascendio la multa a john fashanu?
numerico medid	¿cual es la distancia entre la tierra y el sol?
numerico medid	¿que magnitud tuvo el terremoto que sacudio el norte de japon?
numerico medid	¿quanto mide el everest?
numerico medid	¿que distancia se recorre en el rally granada-dakar?
numerico medid	¿cual es la extension de la selva lacandona?
numerico medid	¿cual es la distancia entre braga y guimarães?
numerico medid	¿cual es la altura del k2?
numerico medid	¿cual es la superficie de la baja sajonia?
numerico medid	¿a que distancia de burgos esta atapuerca?
numerico medid	¿a que distancia de la tierra esta jupiter?
numerico medid	¿que altura tiene el kanchenjunga?
numerico medid	¿que altura tiene la torre eiffel?
numerico medid	¿cuantos kilometros se recorrieron en el tour de 1926?
numerico medid	¿cual es el record del mundo de salto de altura?
numerico edad	¿a que edad murio joseph di mambro?
numerico edad	¿a que edad murio thomas "bp" o'neill?
numerico edad	¿cual era la esperanza de vida en francia en 1991?
numerico edad	¿que edad tenia nick leeson en el momento de ser condenado a la carcel?
numerico edad	¿que edad tenia richard holbrooke en 1995?

10

3. Clasificación utilizando árboles de decisión

Técnicas de clasificación basadas en árboles de decisión:

- Construyen un modelo, hipótesis o representación de la regularidad existente en los datos
- Ventajas respecto a las redes neuronales o las máquinas de vectores de soporte (*Support Vector Machine, SVM*):
 - Son modelos comprensibles porque se pueden expresar de una manera simbólica, en forma de conjunto de condiciones
 - Son eficientes por su característica de algoritmos “voraces” (siempre que quepan todos los ejemplos para aprender en memoria)
 - Hay múltiples implementaciones disponibles
- Desventajas:
 - Son más dependientes del conjunto de ejemplos de aprendizaje

11

3. Clasificación utilizando árboles de decisión

Árbol de decisión:

- Conjunto de condiciones exhaustivas y excluyentes organizadas en una estructura jerárquica
 - Exhaustivo: cada condición ha de cumplirse una de sus opciones ($\text{edad} > 50$ ó $\text{edad} \leq 50$)
 - Excluyente: las particiones del árbol han de ser disjuntas
- La decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas

12

3. Clasificación utilizando árboles de decisión

Ejemplo:

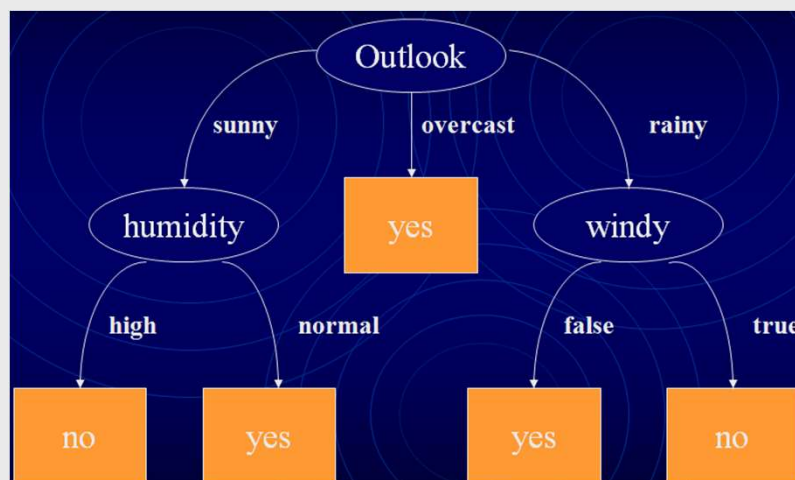
Elección de "Play"

<http://csie.org/~dm/>

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

13

3. Clasificación utilizando árboles de decisión



14

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Explotación de la información, Clasificación y Agrupamiento de Información

Algoritmo para construir AD a partir de datos:

- **Técnica de partición (divide y vencerás):**
 - El espacio de instancias se va partiendo de arriba abajo utilizando cada vez una partición o conjunto de condiciones excluyentes y exhaustivas
 - Una vez elegida la partición, dicha partición no se puede cambiar: *criterio de partición*

15

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Explotación de la información, Clasificación y Agrupamiento de Información

Algoritmo partición (N:nodo; E:conj_ejemplos)

Si todos los ejemplos E son de la misma clase c

Entonces

Asignar clase c al nodo N

Salir // N es hoja

Sino

particiones = generarPosiblesParticiones

MejorPartición = seleccionarMejorParticiónSegún_criterio_partición

Para cada condición i de la partición elegida

Añadir un nodo hijo i a N y asignar los ejemplos consistentes (E_i)

partición (i, E_i) // Llamada recursiva

16

3. Clasificación utilizando árboles de decisión

generarPosiblesParticiones:

■ Tipos de particiones:

- Nominales (x_i): aquellos que tienen un conjunto de posibles valores $\{v_1, v_2, \dots, v_k\}$
 - # Si solo se permiten árboles binarios, la partición será :
 - $(x_i=v_1, x_i \neq v_1), (x_i=v_2, x_i \neq v_2), (x_i=v_3, x_i \neq v_3), \dots$
 - # Caso contrario: $(x_i=v_1, x_i=v_2, \dots, x_i=v_k)$
- Numéricas (x_i): aquellos que tienen un conjunto de posibles valores numéricos y continuos. Las particiones: $(x_i \leq a, x_i > a)$, con a una constante numérica elegida entre un conjunto finito de constantes obtenidas de los ejemplos:
 - # Si x_i presenta los valores $\{0,2 \ 0,3 \ 0,7 \ 0,1 \ 0,8 \ 0,45 \ 0,33 \ 0,1 \ 0,8 \ 0\}$
 - # Se ordenan, eliminan repetidos $\{0 \ 0,1 \ 0,2 \ 0,3 \ 0,33 \ 0,45 \ 0,7 \ 0,8\}$ y se obtienen los valores intermedios $\{0,05 \ 0,15 \ 0,25 \ 0,315 \ 0,39 \ 0,575 \ 0,75\}$ generando particiones binarias:
 - $(x_i \leq 0,05, x_i > 0,05) (x_i \leq 0,15, x_i > 0,15) (x_i \leq 0,25, x_i > 0,25) (x_i \leq 0,315, x_i > 0,315) (x_i \leq 0,39, x_i > 0,39) (x_i \leq 0,575, x_i > 0,575) (x_i \leq 0,75, x_i > 0,75)$

17

3. Clasificación utilizando árboles de decisión

Ejercicio 2:

- Sobre el ejemplo anterior de elección de "Play", a partir de la tabla de ejemplos, obtener las particiones
- Para n atributos y m valores posibles para cada atributo, ¿cuántas particiones se generarían?

18

3. Clasificación utilizando árboles de decisión

seleccionarMejorParticiónSegún_criterio_partición:

- Objetivo: buscar particiones que discriminen más
- Criterio: elegir la partición s con mayor valor $I(s)$

$$I(s) = \sum_{j=1..n} p_j \cdot f(p_j^1, p_j^2, \dots, p_j^c)$$

- n : número de nodos hijos de la partición
- p_j : probabilidad de caer en el nodo j de la partición s
- p_j^l : proporción de elementos de la clase l en el nodo j
- c : número de clases del problema

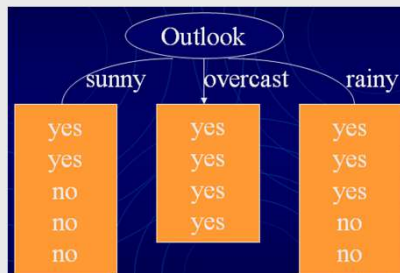
19

3. Clasificación utilizando árboles de decisión


Método basado en entropía C4.5 [Quinlan,93]:

$$I(s) = \sum_{j=1..n} p_j \times f(p_j^1, p_j^2, \dots, p_j^c) = \sum_{j=1..n} \left(p_j \times \sum_{k=1..c} (p_j^k \times \log_2(p_j^k)) \right)$$

$$= \left(\frac{5}{14} \right) \times \left(\frac{2}{5} \times \log_2 \frac{2}{5} + \frac{3}{5} \times \log_2 \frac{3}{5} \right) + \left(\frac{4}{14} \right) \times 0 + \left(\frac{5}{14} \right) \times (-0.971) = -0.693$$



20




Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Ejercicio 3 (evaluación continua):

- Sobre el ejemplo anterior de predicción del tiempo, calcula $I(s)$ para el resto de particiones
- ¿Qué partición quedaría como raíz del árbol de decisión final?

21



Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Ejercicio 4:

- Aplica el algoritmo *partición* para comprobar que se genera el árbol de decisión mostrado anteriormente

22

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Ejercicio 5: Calcula el árbol de decisión

Attributes					Class
Education	Annual Income	Age	Own House	Sex	Credit ranking
College	High	Old	Yes	Male	Good
High school	-----	Middle	Yes	Male	Good
High school	Middle	Young	No	Female	Good
College	High	Old	Yes	Male	Poor
College	High	Old	Yes	Male	Good
College	Middle	Young	No	Female	Good
High school	High	Old	Yes	Male	Poor
College	Middle	Middle	-----	Female	Good
High school	Middle	Young	No	Male	Poor

Explotación de la información, Clasificación y Agrupamiento de Información

4. Clasificación utilizando sistemas de reglas

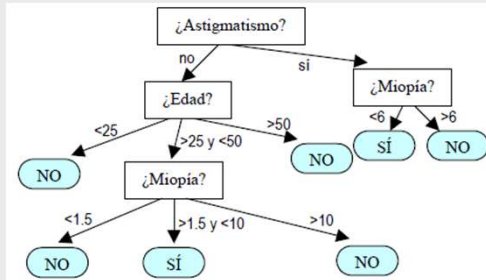
Sistemas de reglas:

- Generalización de los árboles de decisión en el que no se exige exclusión ni exhaustividad en las condiciones de las reglas:
 - Se podría aplicar más de una regla (reglas 1, 3, 5) o ninguna
 - Se agrupan diferentes ramas del árbol en una sola condición: “en otro caso”
- Algoritmo:
 - Se generan reglas sucesivamente, descartándose ejemplos ya cubiertos por las reglas ya obtenidas, y con los ejemplos que quedan se empieza de nuevo

24

4. Clasificación utilizando sistemas de reglas

Sistema para determinar la recomendación de cirugía ocular:



1. SI astig = sí Y miopía > 6 ENTONCES no
2. SI 25 < edad ≤ 50 Y miopía ≤ 6 ENTONCES sí
3. SI edad > 50 ENTONCES no
4. SI edad ≤ 25 ENTONCES no
5. SI miopía > 10 ENTONCES no
6. EN OTRO CASO operación = sí

SI astig = no Y 25 < edad ≤ 50 Y 1.5 < miopía ≤ 10 ENTONCES sí
 SI astig = sí Y miopía ≤ 6 ENTONCES sí
 EN OTRO CASO no

25

4. Clasificación utilizando sistemas de reglas

Algoritmo cobertura(Epos, Eneg:conj_ejemplos)

Reglas = \emptyset

Mientras Epos $\neq \emptyset$ Y NO ParadaReglas // Aprender nueva regla

NuevaRegla = \emptyset

Eneg_Act = Eneg

Mientras Eneg_Act $\neq \emptyset$ Y NO ParadaCondiciones // Aprender nueva
 // condición

Cond = seleccionar una condición según criterio (elimina muchos negativos)

NuevaRegla = NuevaRegla \cup {Cond} // Añadimos la nueva condición a la regla

Eneg_Act = ejemplos negativos consistentes con NuevaRegla

Reglas = Reglas \cup {NuevaRegla}

Epos = Epos – Ejemplos cubiertos por NuevaRegla

Retorna Reglas

26

Table 1.1 The contact lens data.

age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

27

4. Clasificación utilizando sistemas de reglas


Datos de prescripción de lentes *hard*:

age=young	2/8
age=pre-presbyopic	1/8
age=presbyopic	1/8
spectacle prescription=myope	3/12
spectacle prescription=hypermetrope	1/12
astigmatism=no	0/12
astigmatism=yes	4/12
tear production rate=reduced	0/12
tear production rate=normal	4/12


■ Añadimos regla:

- SI astigmatism = yes ENTONCES recommendation = hard

28




4. Clasificación utilizando sistemas de reglas


 Cogemos el resto de ejemplos para refinarla:


age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

age=young	2/4
age=pre-presbyopic	1/4
age=presbyopic	1/4
spectacle prescription=myope	3/6
spectacle prescription=hypermetrope	1/6
tear production rate=reduced	0/6
tear production rate=normal	4/6



4. Clasificación utilizando sistemas de reglas

 Regla refinada:

-  SI astigmatism = yes
 Y tear production rate = normal,
 ENTONCES recommendation = hard

Explotación de la información. Clasificación y Agrupamiento de Información

30

4. Clasificación utilizando sistemas de reglas

■ Seguimos refinando:

Table 4.9 Part of the contact lens data for which astigmatism = yes and tear production rate = normal.

age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	normal	none

31

4. Clasificación utilizando sistemas de reglas

■ Seguimos refinando:

age=young	2/2
age=pre-presbyopic	1/2
age=presbyopic	1/2
spectacle prescription=myope	3/3
spectacle prescription=hypermetrope	1/3

■ Regla refinada:

- SI astigmatism = yes
 - Y tear production rate = normal
 - Y spectacle prescription = myope,
- ENTONCES recommendation = hard

32

4. Clasificación utilizando sistemas de reglas

Ejercicio 6:

- Genera la regla para prescripción de lentes *soft*

Ejercicio 7:

- Genera el árbol de decisión según el algoritmo *partición* visto anteriormente

Ejercicio 8:

- Genera las reglas del ejercicio de elección de “play” visto anteriormente

33

5. Problema del *overfitting*. Sistemas de poda

Sistemas de poda:

- Puede darse situaciones de *overfitting*:

- Que el modelo aprendido se ajuste en exceso a los ejemplos conocidos y funcione mal para los nuevos ejemplos
- Especialmente cuando los ejemplos con los que se aprende contienen “ruido”

- Solución:

- # Obtención de modelos más generales:

- Eliminando condiciones de las ramas del árbol o de algunas reglas

34

5. Problema del *overfitting*. Sistemas de poda

Prepoda:

- Se realiza durante la construcción del árbol o conjunto de reglas
- Se determina el criterio de parada para seguir especializando una rama o regla:
 - N° de ejemplos por nodo, n° de excepciones respecto a la clase mayoritaria, etc.

Pospoda:

- Después de la construcción del árbol o conjunto de reglas
- Se eliminan nodos o reglas en sentido ascendente
- Es menos eficiente que la prepoda

Prepoda + pospoda:

- Algoritmo C4.5 con prepoda por cardinalidad y pospoda más sofisticada

35

6. Part of speech tagging

Objetivo:

- A/AT similar/JJ resolution/NN passed/VBD in/IN the/AT Senate/NN by/IN a/AT vote/NN of/IN 29-5/CD ./.
- Desambiguar:
 - I wouldn't **trust** him.
 - He put money in the family **trust**

Técnicas:

- Basadas en frecuencia de aparición del *tag*.
- Basadas en n-gramas
- Modelos estocásticos

36

6. Part of speech tagging

Basadas en frecuencia de aparición del *tag*:

- $P(t_i | w) = c(w, t_i) / (c(w, t_1) + \dots + c(w, t_k))$
 - $c(w, t_i)$ = número de veces que w/t_i aparece en el corpus
- Éxito: 91% para inglés
- Ejemplo:
 - heat :: noun/89, verb/5

37

6. Part of speech tagging

Transformation-based learning:

- A simple rule-based part of speech tagger. Brill. 1992
- Método:
 1. Etiquetar cada token con el tag más frecuente
 2. Crear reglas que corrijan tags erróneos
 - old_tag new_tag NEXT-TAG tag
 - old_tag new_tag PREV-TAG tag
 - TO IN NEXT-TAG AT
 - NN VB PREV-TAG TO
 3. Contar cuántas correcciones con éxito y fracaso se realizan con cada regla
 4. Seleccionar la mejor regla que maximice: |éxito| - |fracaso|
 5. Si no se alcanza un umbral, ir al paso 2

38

6. Part of speech tagging

Ejercicio 9:

- Sobre el texto etiquetado del ejercicio 2 del módulo 2, obtener reglas que resuelvan errores de etiquetado aplicando la técnica de *Transformation-based learning*.
- Una descripción más detallada de las etiquetas léxicas se puede encontrar en la siguiente transparencia y en <http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html>

39

6. Part of speech tagging

UPenn TreeBank II word tags:

- | | |
|---|---|
| • CC - Coordinating conjunction | • PRP\$ - Possessive pronoun |
| • CD - Cardinal number | • RB - Adverb |
| • DT - Determiner | • RBR - Adverb, comparative |
| • EX - Existential there | • RBS - Adverb, superlative |
| • FW - Foreign word | • RP - Particle |
| • IN - Preposition or subordinating conjunction | • SYM - Symbol |
| • JJ - Adjective | • TO - to |
| • JJR - Adjective, comparative | • UH - Interjection |
| • JJS - Adjective, superlative | • VB - Verb, base form |
| • LS - List item marker | • VBD - Verb, past tense |
| • MD - Modal | • VBG - Verb, gerund or present participle |
| • NN - Noun, singular or mass | • VBN - Verb, past participle |
| • NNS - Noun, plural | • VBP - Verb, non-3rd person singular present |
| • NNP - Proper noun, singular | • VBZ - Verb, 3rd person singular present |
| • NNPS - Proper noun, plural | • WDT - Wh-determiner |
| • PDT - Predeterminer | • WP - Wh-pronoun |
| • POS - Possessive ending | • WPS - Possessive wh-pronoun |
| • PRP - Personal pronoun | • WRB - Wh-adverb |

40

6. Part of speech tagging

Basadas en n-gramas:

■ Corpus de 1000 palabras

■ 1000 uni-gramas

“vice” aparece 50 veces

$$p(\text{vice}) = 50/1000 = 0.05$$

“presidente” aparece 100 veces

$$p(\text{presidente}) = 100/1000 = 0.1$$

“vice presidente”

$$p(\text{vice, presidente}) = 0.05 \times 0.01 = 0.005$$

■ 999 bigramas:

Bigrama “vice – presidente” aparece 40 veces:

$$p(\text{vice-presidente}) = 40/999 = 0.04$$

41

6. Part of speech tagging. Basadas en n-gramas

Pointwise Mutual Information (PMI):

$$\text{PMI}(W_1W_2) = \log \frac{p(W_1W_2)}{p(W_1)p(W_2)} = \log \frac{p(\text{vice presidente})}{p(\text{vice})p(\text{presidente})} = \log \frac{0.04}{0.005} = 2.08$$

- Objetivo: las bigramas con mayor PMI son las más probables

42

6. Part of speech tagging. Basadas en n-gramas

Modelos ocultos de Markov:

- ▣ “La predicción del siguiente estado solo depende del estado actual”

$$p(w_n | w_{n-1}) = \frac{C(w_{n-1}, n)}{C(w_{n-1})}$$

- ▣ Probabilidad de una oración utilizando bigramas:

$$p(w_{0..n}) = \prod_{i=0}^n p(w_i | w_{i-1})$$

43

6. Part of speech tagging. Basadas en n-gramas

Trigramas:

- ▣ La estimación de máxima verosimilitud del trigramma “of the king”:

$$P_{MLE}(\text{KING} | \text{OF THE}) = \frac{\text{count}(\text{OF THE KING})}{\sum_w \text{count}(\text{OF THE } w)} = \frac{\text{count}(\text{OF THE KING})}{\text{count}_{\text{hist}}(\text{OF THE})}$$

44

6. Part of speech tagging

Modelos estocásticos:

- Dada la secuencia de palabras de una oración:
 - $W = w_1, w_2, \dots, w_n$
- Asignar una secuencia de etiquetas:
 - $T = t_1, t_2, \dots, t_n$
- Objetivo:
 - Encontrar T que maximice $P(T|W) = P(W|T) P(T) / P(W) = \alpha$
 $P(W|T) P(T)$
- Forma de cálculo:
 - $P(T) = P(t_1) P(t_2 | t_1) P(t_3 | t_1, t_2) P(t_4 | t_1, t_2, t_3) \dots P(t_n | t_1, t_2, \dots, t_{n-1})$
 $\approx P(t_1) P(t_2 | t_1) P(t_3 | t_2) \dots P(t_n | t_{n-1})$
 - # Utilizando *second order Markov model*: $P(t_i | t_{i-2}, t_{i-1})$;
 - $P(W|T) = P(w_1 | t_1) P(w_2 | t_2) \dots P(w_n | t_n)$

45

6. Part of speech tagging. Modelos estocásticos

Table 1: Statistics to be collected.

notation	counting the number of
C_n	all word tokens w
$C(w)$	occurrences of the word w
$C(w, t)$	occurrences of the word w tagged with t
$C(t)$	occurrences of the tag t
$C(t_1, t_2)$	occurrences of the <i>tag bigram</i> (t_1, t_2) , that is the tag t_1 followed by the tag t_2
$C(t_1, t_2, t_3)$	occurrences of the <i>tag trigram</i> (t_1, t_2, t_3) , that is the tag t_1 followed by t_2 followed by t_3
$C(w_1, t_1, t_2)$	occurrences of the <i>wordtag-tag bigram</i> (w_1, t_1, t_2) , that is the word w_1 tagged with t_1 followed by the tag t_2
$C_m(t)$	different word types tagged with tag t
$C_c(t)$	occurrences of capitalized words tagged with t
$C_m(w_{\text{end-}i}, t)$	different word types ending with the same i letters w and tagged with t

$$P(t_i) = \frac{C(t_i)}{C_n}$$

$$P(t_i | t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

$$P(t_i | t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-1}, t_{i-2})}$$

$$P(t_i | w_{i-1}, t_{i-1}) = \frac{C(w_{i-1}, t_{i-1}, t_i)}{C(w_{i-1}, t_{i-1})}$$

$$P(w_i | t_i) = \frac{C(w_i, t_i)}{C(t_i)}$$

$$P(t_i | w_i) = \frac{C(w_i, t_i)}{C(w_i)}$$

46

Explotación de la información, Clasificación y Agrupamiento de Información

6. Part of speech tagging. Modelos estocásticos

Para ampliar conocimientos:

- “Implementing an efficient part-of-speech tagger”.
Johan Carlberger, Viggo Kann. 24th March 1999
- Google Books: Ngram Viewer
 - <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

47

Explotación de la información, Clasificación y Agrupamiento de Información

6. Part of speech tagging. Modelos estocásticos

Ejercicio 10:

- Dadas las dos siguientes frases:
 - Secretariat/NNP is/VBZ expected/VBN to/TO **race/VB** tomorrow/NN
 - People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT **race/NN** for/IN outer/JJ space/NN
- Y dadas las probabilidades de las bigramas:
 - $P(\text{NN}|\text{TO}) = .021$ $P(\text{race}|\text{NN}) = .00041$
 - $P(\text{VB}|\text{TO}) = .34$ $P(\text{race}|\text{VB}) = .00003$
- Calcular la etiqueta más probable para “race” según el modelo estocástico

48

6. Part of speech tagging. Modelos estocásticos

Añadiendo reglas:

- Detección de nombres propios si la palabra empieza por mayúscula:

$$P_c(w, t) = \begin{cases} \gamma_1 & \text{if } t \text{ is not proper-noun tag and } w \text{ is capitalized,} \\ \gamma_2 & \text{if } t \text{ is proper-noun tag and } w \text{ is not capitalized,} \\ 1 & \text{otherwise.} \end{cases}$$

- $\gamma_1 = 0.028$ and $\gamma_2 = 0.044$
- En el caso de palabras desconocidas: $\gamma_1 = 0.020$ $\gamma_2 = 0.048$

$$T(w_{1..n}) = \operatorname{argmax}_{t_{1..n}} \prod_{i=1}^n P_{int}(t_i | t_{i-2}, t_{i-1}) P(w_i | t_i) P_c(w_i, t_i).$$

49

6. Part of speech tagging. Modelos estocásticos

Etiquetando palabras desconocidas:

- Hay que estimar $P_m(w | t)$ en lugar de $P(w | t)$:
éxito del 45.5% en etiquetado de palabras desc.

$$P_m(w | t) = \frac{C_m(t)}{\sum_{\tau \in \text{tag set}} C_m(\tau)}$$

- Se puede añadir frecuencias de terminaciones (L máximo de 5, éxito del 88.7%):

$$P_e(w | t) = \sum_{i=0}^L \alpha_i \cdot \frac{C(w_{\text{end-i}}, t)}{\sum_{\tau \in \text{tag set}} C(w_{\text{end-i}}, \tau)}$$

50

Explotación de la información, Clasificación y Agrupamiento de Información

7. Sistemas de agrupamiento de Información

51

Agrupamiento (*clustering*):

- Separar en grupos basándose en las similitudes o relaciones existentes
- Diferencias con la clasificación automática:
 - Los grupos o categorías no están necesariamente predefinidos
 - Se pueden asignar uno o varios grupos
- Aplicaciones:
 - Recuperación de información: organizar los resultados
 - Facilitar la navegación por una colección de documentos
 - Creación de directorios Web (*Yahoo*)

Explotación de la información, Clasificación y Agrupamiento de Información

7. Sistemas de agrupamiento de Información

52

Agrupamiento en la RI:

- **Objetivo:**
 - Particionar una colección de documentos D en k subconjuntos o clusters D_1, D_2, \dots, D_k , de tal forma que se minimice la distancia intracluster o se maximice la semejanza intracluster:
 - # Utilizando el modelo vectorial:
 - Un clúster sería un *centroide* de los documentos
 - Objetivo:
 - Minimizar $\sum_i \sum_{d \in D_i} \text{distancia}(d, \bar{D}_i)$ o maximizar $\sum_i \sum_{d \in D_i} \text{semejanza}(d, \bar{D}_i)$
- **Hipótesis de agrupamiento:**
 - Los documentos fuertemente asociados tienden a ser relevantes para la misma consulta
 - Si un usuario está interesado en un doc de un grupo, también es probable que lo esté en los demás miembros del grupo

8. Sistemas de agrupamiento de Información en la RI

Tipos de agrupamiento en la RI:

■ *Pre-retrieval document clustering:*

- Se realiza en fase de indexación
- Se elige un representante del grupo que sería con el que se compara la query (los restantes docs del grupo no se comparan)
- Problema: creación de grupos estáticos en un entorno tan dinámico como es la Web

■ *Post-retrieval document clustering:*

- Se realiza en fase de presentación de resultados de la fase de búsqueda
- Se agrupan los documentos devueltos por el motor de búsqueda
- Problema: eficiencia del proceso en tiempo de búsqueda

53

8. Sistemas de agrupamiento de Información en la RI

Fases en el agrupamiento en la RI:

- Selección/extracción de características: representación de objetos
- Cálculo de la similitud entre objetos: medidas de distancia
- Clustering o agrupamiento

54

8. Sistemas de agrupamiento de Información en la RI

Técnicas de agrupamiento:

- **No exclusivas:** un doc puede pertenecer a varios grupos
- **Exclusivas:** un doc solo pertenece a un grupo
 - **Extrínsecas:**
 - # Cuando los grupos están predefinidos y se tienen objetos que ya están agrupados en dichos clusters, los cuales son utilizados por el algoritmo para aprender a agrupar el resto de objetos
 - **Intrínsecas:**
 - # Los grupos se crean a partir de las características propias de los objetos sin conocer previamente los grupos
 - # Tipos:
 - **Jerárquicas:** los grupos se consiguen mediante la separación o unión de grupos de documentos generando una estructura en árbol con grupos anidados
 - **Particionales:** se llega a un agrupamiento que optimiza un criterio predefinido o función objetivo, creando una estructura plana, sin grupos anidados

55

9. Sistemas de agrupamiento de información particionales

Técnicas de agrupamiento *particionales* (*k-clustering, k-means, k-medoids*):

- **Algoritmo:**
 - Se determina a priori el *número de grupos*:
 - # Se cogen los *k* primeros objetos, o
 - # Los *k* objetos más alejados entre sí, o
 - # *k* objetos aleatoriamente
 - Iterativamente se van asignando docs a estas particiones
 - Los docs se reasignan de acuerdo a una *función objetivo*
 - El proceso se repite hasta que se consigue un *criterio de terminación*
- **Variaciones de los clusters:**
 - Juntar grupos cuando la distancia entre sus centroides esté por debajo de un umbral
 - Dividir grupos cuando su varianza esté por encima de un umbral

56

9. Sistemas de agrupamiento de información particionales

Función objetivo:

■ Internas: miden similitud *intra-cluster*:

- Maximizar la suma de los promedios de las similitudes existentes entre los pares de docs asignados a cada cluster, teniendo en cuenta el tamaño de cada uno:

k: n° de clusters; n: n° elementos de cada cluster; sim(d, e): función de similitud p.ej. el coseno

$$\max I_1 = \sum_{r=1}^k n_r \times \left(\frac{1}{n_r^2} \times \sum_{d_i, d_j \in S_r} \text{sim}(d_i, d_j) \right)$$

■ Externas: miden distancia *inter-cluster*:

- Minimizar similitud entre centroide de cada cluster y el centroide de la colección completa

$$\min E_1 = \sum_{r=1}^k n_r \times \text{sim}(C_r, C)$$

57

9. Sistemas de agrupamiento de información particionales

k-mean:

- Generar los *k* clusters iniciales con sus docs
- Inicializar los centroides de cada cluster
- Mientras sea posible realizar más mejoras

■ Para cada documento *d*

- # Encontrar el cluster *c* cuyo centroide es más similar a *d*
- # Asignar *d* al cluster *c*

■ Para cada cluster *c*

- # Recalcular el centroide de *c* según los documentos asignados a *c*

58

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Ejemplo de aplicación de *k-mean* (University of South Carolina Upstate, Angelina Tzacheva):

- Supongamos:
 - Los siguientes 8 vectores: A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9)
 - k=3
 - Clusters iniciales: A1(2, 10), A4(5, 8), A7(1, 2)
 - Distancia entre dos vectores $a=(x1, y1)$ y $b=(x2, y2)$:

$$\# \rho(a, b) = |x2 - x1| + |y2 - y1|$$
 - Centroide de un grupo n de vectores: vector con el resultado de la media de los n vectores. Cada componente del vector centroide será la media aritmética de las casillas de todos los vectores

59

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

		Cluster 1 (2, 10)	Cluster 2 (5, 8)	Cluster 3 (1, 2)	
	Vector	Dist Clust 1	Dist Clust 2	Dist Clust 3	Cluster
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

60

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

Clusters conseguidos:

Cluster 1	Cluster 2	Cluster 3
(2, 10)	(8, 4)	(2, 5)
	(5, 8)	(1, 2)
	(7, 5)	
	(6, 4)	
	(4, 9)	

61

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

Recálculo de los centroides:

- Cluster 1: (2, 10)
- Cluster 2: $((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$
- Cluster 3: $((2+1)/2, (5+2)/2) = (1.5, 3.5)$

62

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Ejercicio 9 (evaluación continua):

- ▣ Calcular las dos siguientes iteraciones del algoritmo *k-means*

63

Explotación de la información, Clasificación y Agrupamiento de Información

10. Sistemas de agrupamiento de información jerárquicos

Dendograma:

- ▣ Estructura en árbol de clusters o grupos
- ▣ Las hojas son grupos que contienen un único documento
- ▣ La raíz es un único grupo con todos los elementos de la colección
- ▣ Los niveles intermedios son las posibles configuraciones de clusters

64

10. Sistemas de agrupamiento de información jerárquicos

Tipos de sistemas jerárquicos:

■ Aglomerativos:

- Se comienza con los objetos o individuos de modo individual
- Luego se van agrupando de modo que los primeros en hacerlo son los más similares
- Al final, todos los subgrupos se unen en un único cluster

■ Divisivos:

- Se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén

65

10. Sistemas de agrupamiento de información jerárquicos

Sistemas jerárquicos aglomerativos. Algoritmo:

- Empezar con N clusters (el número inicial de elementos) y una matriz $N \times N$ simétrica de distancias o similitudes. $D = [d_{ik}]_{ik}$.
- Dentro de D , buscar aquella entre los clusters U y V (más próximos, más distantes o en media más próximos) que sea la menor entre todas, d_{uv}
- Juntar U y V en uno solo. Actualizar D :
 - Borrando las filas y columnas de los clusters U y V
 - Formando la fila y columna de las distancias del nuevo cluster (UV) al resto de clusters
- Repetir los pasos (2) y (3) un total de $(N - 1)$ veces

66

10. Sist. agrupamiento información jerárquicos aglomerativos

Ejemplo (Univ. Carlos III, J.M. Marin):

■ Primera iteración (5 objetos):

$$D = [d_{ik}]_{ik} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{matrix}$$

■ La menor distancia (2) hace que se unan 3 y 5

■ Distancia entre el cluster (35) y los objetos 1, 2, 4

$$d_{(35),1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$

$$d_{(35),2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$

$$d_{(35),4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

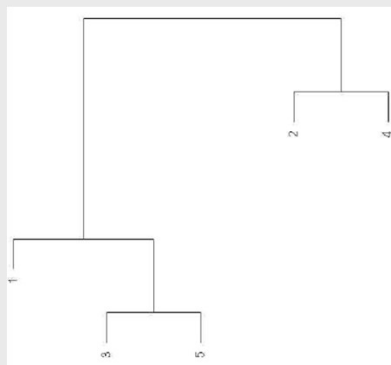
$$\begin{matrix} & \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} \\ \begin{matrix} (35) \\ 1 \\ 2 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

67

10. Sist. agrupamiento información jerárquicos aglomerativos

Ejercicio 11:

■ Calcular las siguientes iteraciones del algoritmo para calcular el dendograma completo



68