Université M[ed] Boudiaf -M'sila
Faculté des Mathématiques et d'Informatique
Département d'Informatique
Module : Data Mining et Recherche d'Information
Master 2/ S.I. Avancés

Durée : 1h 30

Examen Semestriel

Exercice 1 (2 points)

En utilisant Google, vous tapez le mot « souvenir » puis le mot « qouvenir ». Quelle est la réponse de Google pour le premier et le 2ème mot ? Expliquez.

Exercice 2 (4 points)

Etant données les valeurs ***tf*** et ***df*** du tableau ci-contre, calculer pour chaque document, le ***tf*idf*** des termes suivants: **car, auto, assurance, bon**.

Il y a au total 200 000 documents. On utilisera le logarithme base 10.

Terme	df	Doc1-tf	Doc2-tf	Doc3-tf
Car	200	1	100	10
auto	20	1	10	1
assurance	2 000	100	10	1
bon	20 000	100	1 000	10

Exercice 3 (4 points)

Considérer la matrice de dissimilarité suivante P.
Déterminer les dendrogrammes résultants de l'application du « **single link algorithm** », puis du « **complete link algorithm** » sur P et commentez.

a	b	c	d	e
0	4	9	6	5
4	0	3	8	7
9	3	0	3	2
6	8	3	0	1
5	7	2	1	0

Exercice 4 (5 points)

Soit le tableau de données suivant.
Les valeurs possibles des attributs du tableau sont :
Humidité : normale, haute
Climat : soleil, couvert, pluie
Température : chaud, bon, frais

La colonne « Golf » donne les classes :
Jouer, **NePasJouer.**

En utilisant l'entropie comme critère de sélection, déterminer l'attribut qui sera la racine de l'arbre de décision.
N.B. On pourra utiliser la formule :
$\log_2(x) = \log_{10}(x) / \log_{10}(2)$.

N°	Humidité	Climat	Température	Golf
1	haute	soleil	chaud	NePasJouer
2	haute	soleil	chaud	NePasJouer
3	haute	couvert	chaud	Jouer
4	haute	pluie	bon	Jouer
5	normale	pluie	frais	Jouer
6	normale	pluie	frais	NePasJouer
7	normale	couvert	frais	Jouer
8	haute	soleil	bon	NePasJouer
9	normale	soleil	frais	Jouer
10	normale	pluie	bon	Jouer
11	normale	soleil	bon	Jouer
12	haute	couvert	on	Jouer
13	normale	couvert	chaud	Jouer
14	haute	pluie	bon	NePasJouer

Exercice 5 (5 points)

Soit la table suivante des transactions d'un magasin.
En utilisant l'algorithme **Apriori**, déterminer l'ensemble des itemset fréquents et extraire les différentes règles possibles.
(Support Minimum=10%).

Transaction ID	Items
001	A, C, D
002	B, C, E
003	A, B, C, E
004	B, E

Bonne réussite
T. Mehenni

DATA MINING ET RECHERCHE D'INFORMATION **Responsable du cours :**
CORRECTION DE L'EXAMEN **T. Mehenni**

Exercice 1 (2 points)

Le moteur Google affichera les résultats de recherche du mot « souvenir ».
Pour le mot « qouvenir », Google affichera un message que le mot correct est peut être « souvenir », et donne les résultats du mot « souvenir » car il ne peut pas trouver des résultats correspondants au mot « qouvenir ».
Google est muni de cette fonctionnalité qui propose à l'utilisateur des mots proches du mot tapé, quand celui-ci est incorrect, mal orthographié ou non familier. Pour le mot « qouvenir », Google estime qu'il y a eu peut être une erreur de frappe et que l'utilisateur au lieu de taper le mot « souvenir », a tapé le mot « qouvenir » car les deux lettres « s » et « q » sont proches sur le clavier.

Exercice 2 (4 points)

On a tf*idf=tf*log(d/df).

- Mot « **car** » : idf=log(200 000/200)=3 …….(4* 0.25 pt)
  - Doc1 : tf*idf = 1 * 3 = 3
  - Doc2 : tf*idf = 100 * 3 = 300
  - Doc3 : tf*idf = 10 * 3 = 30
- Mot « **auto** » : idf=log(200 000/20)=4 …….(4* 0.25 pt)
  - Doc1 : tf*idf = 1 * 4 = 4
  - Doc2 : tf*idf = 10 * 4 = 40
  - Doc3 : tf*idf = 1* 4 = 4
- Mot « **assurance** » : idf=log(200 000/2000)=2 …….(4* 0.25 pt)
  - Doc1 : tf*idf = 100 * 2 = 200
  - Doc2 : tf*idf = 10 * 2 = 20
  - Doc3 : tf*idf = 1 * 2 = 2
- Mot « **bon** » : idf=log(200 000/20 000)=1 …….(4* 0.25 pt)
  - Doc1 : tf*idf = 100 * 1 =100
  - Doc2 : tf*idf = 1000 * 1 = 1000
  - Doc3 : tf*idf = 10 * 1 = 10

Exercice 3 (4 points)

- **Simple link algorithm :** …….(4* 0.25 pt)

	a	**b**	**c**	**d**	**e**
a	0				
b	4	0			
c	9	3	0		
d	6	8	3	0	
e	5	7	2	**1**	0

grouper (a, b)

	a	**b**	**c**	**de**
a	0			
b	4	0		
c	9	3	0	
de	5	7	**2**	0

grouper (c, de)

	a	**b**	**Cde**
a	0		
b	4	0	
cde	5	**3**	0

grouper (b, cde)

	a	**bcde**
a	0	
bcde	**4**	0

grouper(a, bcde)

- **Complete link algorithm :** …….(4 * 0.25 pt)

	a	**b**	**c**	**d**	**e**
a	0				
b	4	0			
c	9	3	0		
d	6	8	3	0	
e	5	7	2	**1**	0

grouper (a, b)

	a	**b**	**c**	**de**
a	0			
b	4	0		
c	9	3	0	
de	6	8	**3**	0

grouper (c, de)

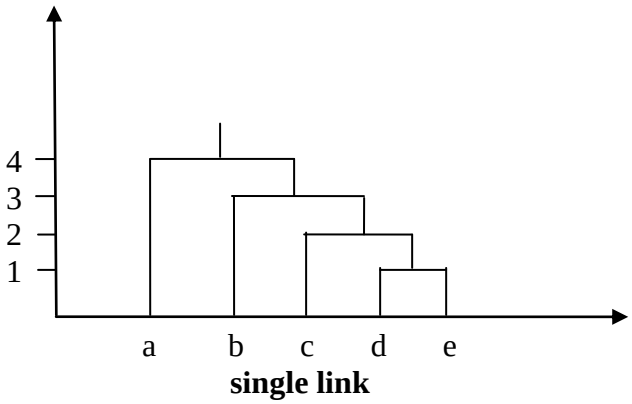
	a	**b**	**cde**
a	0		
b	**4**	0	
cde	9	8	0

grouper (a, b)

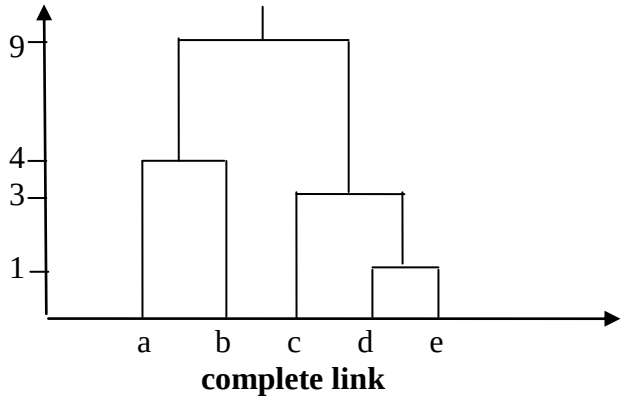
	ab	**cde**
ab	0	
cde	**9**	0

grouper(a, bcde)

Les dendrogrammes : …….(2 pts)



single link



complete link

Exercice 4 (5 points)

Entropie =-p(c_1) log(p(c_1)) - p(c_1) log(p(c_1))
Gain=Entropie-n1/n [p(c_1) log(p(c_1))] – n2/n [p(c_1) log(p(c_1))]

Calcul de l'entropie des données initiales :
C1 : 9 lignes concernant la classe « Jouer »
C2 : 5 lignes concernant la classe « NePasJouer »
Total=14
Entropie = -9/14 log(9/14) - 5/14 log(5/14) = 0.123 + 0.159 = **0.282** ……..…….(1 pt)

Calcul du gain de l'attribut **Humidité** :
Répartition des classes selon les valeurs de l'attribut : …….(0.5 pt)

	Jouer	NePasJouer	n_i/n
P1 (haute)	3	4	7/14
P2 (normale)	6	1	7/14

Entropie(P1)= - 3/7 log(3/7) - 4/7 log(4/7) = 0.157 + 0.138 = 0.259
Entropie(P2)= - 6/7 log(6/7) - 1/7 log(1/7) = 0.057 + 0.120 = 0.177

Gain(Humidité)= 0.282 – 7/14 * 0.259 – 7/14 * 0.177 = **0.046** …….…….…….(0.5 pt)

Calcul du gain de l'attribut **Climat** :
Répartition des classes selon les valeurs de l'attribut : …….(0.5 pt)

	Jouer	NePasJouer	n_i/n
P1 (soleil)	2	3	5/14
P2 (couvert	4	0	4/14
P3 (pluie)	3	2	5/14

Entropie(P1)= - 2/5 log(2/5) - 3/5 log(3/5) = 0.159 + 0.133 = 0.292
Entropie(P2)= - 4/4 log(4/4) = 0
Entropie(P3)= - 3/5 log(3/5) - 2/5 log(2/5) = 0.133 + 0.159 = 0.292

Gain(Climat)= 0.282 – 5/14 * 0.292 – 5/14 * 0.292 = **0.073** …….(0.5 pt)

Calcul du gain de l'attribut **Température** :
Répartition des classes selon les valeurs de l'attribut : …….(0.5 pt)

	Jouer	NePasJouer	n_i/n
P1 (chaud)	2	2	4/14
P2 (bon)	4	2	6/14
P3(frais)	3	1	4/14

Entropie(P1)= - 2/4 log(2/4) - 2/4 log(2/4) = 0.301
Entropie(P2)= - 4/6 log(4/6) - 2/6 log(2/6) = 0.117 + 0.159 = 0.276
Entropie(P3)= - 3/4 log(3/4) - 1/4 log(1/4) = 0.093 + 0.150 = 0.243

Gain(Température)= 0.282 – 4/14 * 0.301 – 6/14 * 0.276 – 4/14 * 0.243 = **0.001** …….(0.5 pt)

L'attribut choisi pour la racine de l'arbre de décision est donc « **Climat** ». …….(1 pt)

Exercice 5 (5 points)

1-Itemset : …….....(0.5 pt)

Item	Fréquence	Support
A	2	2/5=0.5 (50%)
B	3	3/5=0.6 (60%)
C	3	0.6 (60%)
D	1	0.25 (25%)
E	3	0.6 (60%)

On garde tous les 1-itemset car ils ont un support > 10%.

2-Itemset : ……..…….…….(0.5 pt)

	A	B	C	D	E
A	X	1	2	1	1
B	X	X	2	0	2
C	X	X	X	1	2
D	X	X	X	X	0

On garde les 2-itemset suivants : AB, AC,AD, AE, BC, BE, CD et CE (8 2-itemset).

3-Itemset : ……..…….(1 pt)

	AB	AC	AD	AE	BC	BE	CD	CE
AB	X	1	0	1	Déjà vu	Déjà vu	X	X
AC	X	X	1	1	Déjà vu	X	Déjà vu	Déjà vu
AD	X	X	X	0	X	X	Déjà vu	X
AE	X	X	X	X	X	Déjà vu	X	Déjà vu
BC	X	X	X	X	X	2	0	Déjà vu
BE	X	X	X	X	X	X	X	Déjà vu
CD	X	X	X	X	X	X	X	0

On garde les 5 3-itemset suivants : ABC, ABE, ACD, ACE, BCE.

4- Itemset : …….(1 pt)

	ABC	ABE	ACD	ACE	BCE
ABC	X	1	0	Déjà vu	Déjà vu
ABE	X	X	X	Déjà vu	Déjà vu
ACD	X	X	X	0	X
ACE	X	X	X	X	Déjà vu

Un seul 4-itemset, ABCE.

Générations des règles :

1- Règles générées à partir des 2-itemset : il y a 8 * 2 = 16 règles au total. …….(0.5 pt)

Si	A	Alors	B	Si	B	Alors	A
Si	A	Alors	C	Si	C	Alors	A
Si	A	Alors	D	Si	D	Alors	A
Si	A	Alors	E	Si	E	Alors	A
Si	B	Alors	C	Si	C	Alors	B
Si	B	Alors	E	Si	E	Alors	B
Si	C	Alors	D	Si	D	Alors	C
Si	C	Alors	E	Si	E	Alors	C

1- Règles générées à partir des 3-itemset : il y a 3 *5 *2 = 30 règles au total. ……..….(0.75 pt)

Si	A	Alors	BC	Si	BC	Alors	A
Si	A	Alors	BE	Si	BE	Alors	A
Si	A	Alors	CD	Si	CD	Alors	A
Si	A	Alors	CE	Si	CE	Alors	A
Si	B	Alors	CE	Si	CE	Alors	B
Si	B	Alors	AC	Si	AC	Alors	B
Si	B	Alors	AE	Si	AE	Alors	B
Si	C	Alors	AD	Si	AD	Alors	C
Si	C	Alors	AE	Si	AE	Alors	C
Si	C	Alors	BE	Si	BE	Alors	C
Si	C	Alors	AB	Si	AB	Alors	C
Si	E	Alors	AB	Si	AB	Alors	E
Si	D	Alors	AC	Si	AC	Alors	D
Si	E	Alors	AC	Si	AC	Alors	E
Si	E	Alors	BC	Si	BC	Alors	E

2- Règles générées à partir du 4-itemset : il y a 14 règles au total. …….(0.75 pt)

Si	A	Alors	BCE	Si	BCE	Alors	A
Si	B	Alors	ACE	Si	ACE	Alors	B
Si	C	Alors	ABE	Si	ABE	Alors	C
Si	E	Alors	ABC	Si	ABC	Alors	E
Si	AB	Alors	CE	Si	CE	Alors	AB
Si	AC	Alors	BE	Si	BE	Alors	AC
Si	AE	Alors	BC	Si	BC	Alors	AE