Université Mohamed Boudiaf M'sila
Faculté MI, Dépt Informatique

M1 IA
M'sila, Le 05/06/2022
Durée : 90 Minutes

Examen Final (Corrigé)

Machine Learning et Data Mining

Exercice 1 : (10 Points)

L'ensemble de données est composé de 20 lignes (variétés de champignons). Les variétés peuvent être répétées (on peut imaginer les avoir trouvé dans la forêt par exemple). Chaque variété peut avoir les caractéristiques suivantes :
Cap-surface: fibrous , scaly, smooth . On les code: C1, C2, C3
Bruises: bruises, no. On les code: B1, B2
Gill-size: broad, narrow. On les code: G1, G2
Habitat:, grasses, leaves, paths, waste, woods. On les code: H1, H2, H3, H4, H5
Poisonousness: edible, poisonous. On les code: P1, P2

1- On commence par trouver tous les 1-itemsets (itemsets de taille 1) et leurs supports. **(2pts)**

Les 1-itemsets fréquents (ayant un support minimal égal à 25 % (0.25) sont : C1, C2, C3, B1, B2, G1, G2, H1, H5, P1, P2.

Code	Valeur	Fréquence	Support
C1	fibrous	6	0.30
C2	scaly	9	0.45
C3	smooth	5	0.25
B1	bruises	7	0.35
B2	no	13	0.65
G1	broad	13	0.65
G2	narrow	7	0.35
H1	grasses	5	0.25
H2	leaves	4	0.20
H3	paths	3	0.15
H4	waste	1	0.05
H5	woods	7	0.35
P1	edible	10	0.50
P2	poisonous	10	0.50

Cap-surface	Bruises	Gill-size	Habitat	Poisonousness
scaly	bruises	broad	waste	edible
smooth	no	narrow	woods	poisonous
fibrous	no	broad	grasses	edible
scaly	bruises	broad	woods	edible
scaly	no	narrow	leaves	poisonous
scaly	bruises	broad	paths	edible
smooth	no	broad	leaves	edible
scaly	no	broad	woods	poisonous
scaly	no	narrow	woods	poisonous
smooth	no	broad	leaves	edible
fibrous	no	broad	paths	poisonous
fibrous	bruises	broad	woods	edible
smooth	bruises	narrow	grasses	poisonous
fibrous	no	broad	paths	poisonous
smooth	bruises	narrow	grasses	poisonous
scaly	no	narrow	leaves	poisonous
scaly	no	narrow	woods	poisonous
fibrous	no	broad	grasses	edible
scaly	bruises	broad	woods	edible
fibrous	no	broad	grasses	edible

Sur la base des 1-itemsets fréquents, on génère les 2-itemsets. **(2.5pts)**

		C1	C2	C3	B1	B2	G1	G2	H1	H5	P1	P2
1		fibrous	scaly	smooth	bruises	no	broad	narrow	grasses	woods	edible	poisonous
C1	fibrous	0	0	0	1	**5**	**6**	0	3	1	4	2
C2	scaly		0	0	4	**5**	**5**	4	0	**5**	4	**5**
C3	smooth			0	2	3	2	3	2	1	2	3
B1	bruises				0	0	**5**	2	2	3	**5**	2
B2	no					0	**8**	**5**	3	4	**5**	**8**
G1	broad						0	0	3	4	**10**	3
G2	narrow							0	2	3	0	**7**
H1	grasses								0	0	3	2
H5	woods									0	3	4
P1	edible										0	0
P2	poisonous											0

Les 2-itemsets fréquents sont : (C1, B2), (C1, G1), (C2, B2), (C2, G1), (C2, H5), (C2, P2), (B1, G1), (B1, P1), (B2, G1), (B2, G2), (B2, P1), (B2, P2), (G1,P1), (G2, P2)

Sur la base des 2-itemsets fréquents, on génère les 3-itemsets. **(2.5pt)**

		C1	C1	C2	C2	C2	C2	B1	B1	B2	B2	B2	B2	G1	G2
		B2	G1	B2	G1	H5	P2	G1	P1	G1	G2	P1	P2	P1	P2
C1	B2	0	5	x	x	x	x	x	x	x	0	3	2	x	x
C1	G1		0	x	x	x	x	1	x	x	x	x	x	4	x
C2	B2			0	1	3	5	x	x	x	4	0	x	x	x
C2	G1				0	3	4	4	x	1	x	x	x	4	x
C2	H5					0	3	x	x	x	x	x	x	x	x
C2	P2						0	x	x	x	x	x	x	x	x
B1	G1							0	5	x	x	x	x	x	x
B1	P1								0	x	x	x	x	x	x
B2	G1									0	x	5	3	x	x
B2	G2										0	0	x	x	5
B2	P1											0	x	x	x
B2	P2												0	x	x
G1	P1													0	x
G2	P2														0

Les 3-itemsets fréquents sont : (C1,B2,G1), (C2,B2,P2), (B1,G1,P1), (B2,G1,P1), (B2,G2,P2)

Sur la base des 3-itemsets fréquents, on génère les 4-itemsets. **(1pt)**

			C1	C1	B1	B2	B2
			B2	G1	P1	P1	P2
			G1	P2	P1	P1	P2
C1	B2	G1	0	2	3	x	x
C2	B2	P2		0	x	x	x
B1	G1	P1			0	x	x
B2	G1	P1				0	x
B2	G2	P2					0

Il n'y a aucun 4-itemset fréquent.

On génère les règles d'association à partir des ensembles 2-itemset et 3-itemset. **(2pts)**

Items	Règle	Support	Confiance	Items	Règle	Support	Confiance
(C1, B2)	C1 → B2	5/20=25%	5/6=83%	(C1,B2,G1)	C1 → B2,G1		
	B2 → C1	25%	5/13		B2,G1 → C1		
(C1, G1)	C1 → G1	6/20	6/6=100%		B2 → C1,G1		
	G1 → C1	6/20	6/13		C1,G1 → B2		
(C2, B2)	C2 → B2				G1 → C1,B2		
	B2 → C2				C1,B2 → G1		
(C2, G1)	C2 → G1			(C2,B2,P2)			
	G1 → C2						
(C2, H5)	C2 → H5						
	H5 → C2						
(C2, P2)	C2 → P2						
	P2 → C2						
(B1, G1)	B1 → G1			(B1,G1,P1)			
	G1 → B1						
(B1, P1)	B1 → P1						
	P1 → B1						
(B2, G1)	B2 → G1						
	G1 → B2						
(B2, G2)	B2 → G2			(B2,G1,P1)			
	G2 → B2						
(B2, P1)	B2 → P1						
	P1 → B2						
(B2, P2)	B2 → P2						
	P2 → B2						
(G1,P1)	G1 → P1			(B2,G2,P2)			
	P1 → G1						
(G2, P2)	G2 → P2						
	P2 → G2						

On retient seulement les règles d'association ayant support minimal=25% et confiance minimale =90%.

Exercice 2 (6 points)

L'observation de quatre variables binaires sur quatre individus différents a conduit aux résultats suivants (tableau ci-contre). Donner les deux clusters constituant ces quatre individus en appliquant l'algorithme KMeans (utiliser l'indice de Jaccard).

Individu	Variable 1	Variable 2	Variable 3	Variable 4
A	1	1	0	1
B	1	0	0	0
C	0	1	1	0
D	0	1	1	1

Indice (ou coefficient) de Jaccard : $d(i,j)=\frac{b+c}{a+b+c}$

On choisit aléatoirement deux centres, soient A et C

On calcule les différentes distances entre les points.

Dist(B,A)=(0+2)/(1+0+2)=2/3=0.66 **(1 Pt)**

		A	
		1	0
B	1	1	0
	0	2	1

Dist(B,C)=(1+2)/(0+1+2)=3/3=1.0

D'où A est affecté au cluster de centre B **(1 Pt)**

		C	
		1	0
B	1	0	1
	0	2	1

Dist(D,A)=(1+1)/(2+1+1)=2/4=0.50 **(1 Pt)**

		A	
		1	0
D	1	2	1
	0	1	0

Dist(D,C)=(1+0)/(2+1+0)=1/3=0.33

D'où D est affecté au cluster de centre C **(1 Pt)**

		C	
		1	0
D	1	2	1
	0	0	1

Les deux clusters sont : C1={A, B}, C2={C, D} **(2 Pts)**

Exercice 3 (4 points)

Les deux caractéristiques des plantes: couleur des feuilles (jaune, vert, rouge) et largeur des feuilles (petite, grande).

Soient les données des plantes A et B suivantes : Plante A : feuilles rouges et petites. Plante B : feuilles jaunes et grandes.

On a en tout cinq variables binaires (trois pour la couleur et deux pour la taille des feuilles).

D'où on obtient le tableau suivant : **(3 Pts)**

	Jaune	Vert	Rouge	Petite	Grande
Plante A	0	0	1	1	0
Plante B	1	0	0	0	1

Dist(A,B)=(2+2)/(0+2+2)=4/4=1.0 **(1 Pt)**

		B	
		1	0
A	1	0	2
	0	2	1

Rédigé par
Dr. Tahar Mehenni