Data Mining & Data Warehousing, 10 juin 2009, PARTIE CAHIER FERMÉ

Nom et prénom
Année

Durée : 1 heure et 15 minutes. Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Question 1 Occam's Razor.

.../5

Question 2 Prepruning.

.../5

Question 3 $F_{k-1} \times F_{k-1}$ method.

.../5

Question 4 Single Link.

.../5

Question 5 Bisecting K-means.

. . . /5

Question 6 Tree replication problem.

. . . /5

Question 7 Closed frequent itemset.

. . . /5

Question 8 Silhouette coefficient.

. . . /5

Nom et prénom
Année

Durée : 1 heure et 45 minutes.

Question 9 6 Le magazine *Top Sport* met en ligne un vaste nombre d'articles sur tous les sports. Un adepte de tennis s'intéresse aux articles sur le tennis, et rien qu'aux articles sur le tennis. Au lieu de chercher ces articles "à la main", il envisage de construire un classificateur pour classer les articles en deux classes : ceux qui traitent du tennis (Tennis="oui") et les autres (Tennis="non"). Pour ce faire, il dispose d'une table qui enregistre le nombre d'occurrences de certains mots clé dans chaque article. Par exemple, l'article `doc1.pdf` contient 12 fois le mot "Saive", 0 fois le mot "Henin", etc.; cet article ne relève pas du tennis.

Article	#Saive	#Henin	#Wimbledon	#Beijing	...	Tennis
doc1.pdf	12	0	0	5		non
doc2.pdf	0	17	1	3		oui
			⋮			⋮

Les matrices de confusion se présentent comme suit:

		Predicted	
		Tennis=oui	Tennis=non
Observed	Tennis=oui	*TP*	*FN*
	Tennis=non	*FP*	*TN*

,

avec TP, TN, FP, TN des entiers ≥ 0. À partir d'une telle matrice, deux mesures de qualité sont calculées:

$$\pi = \frac{TP}{TP + FP} \qquad \text{et} \qquad \rho = \frac{TP}{TP + FN}$$

Sur un ensemble de test, on obtient les valeurs de π et ρ suivantes pour trois programmes de classification:

	π	ρ
Naive Bayes	0.80	0.80
C4.5	0.90	0.20
MultilayerPerceptron	0.20	0.90

Si vous deviez choisir un modèle de prédiction à partir de ces chiffres, quel serait ce choix (cocher une case) ?

□ Naive Bayes □ C4.5 □ MultilayerPerceptron

Justifiez votre choix en détail.

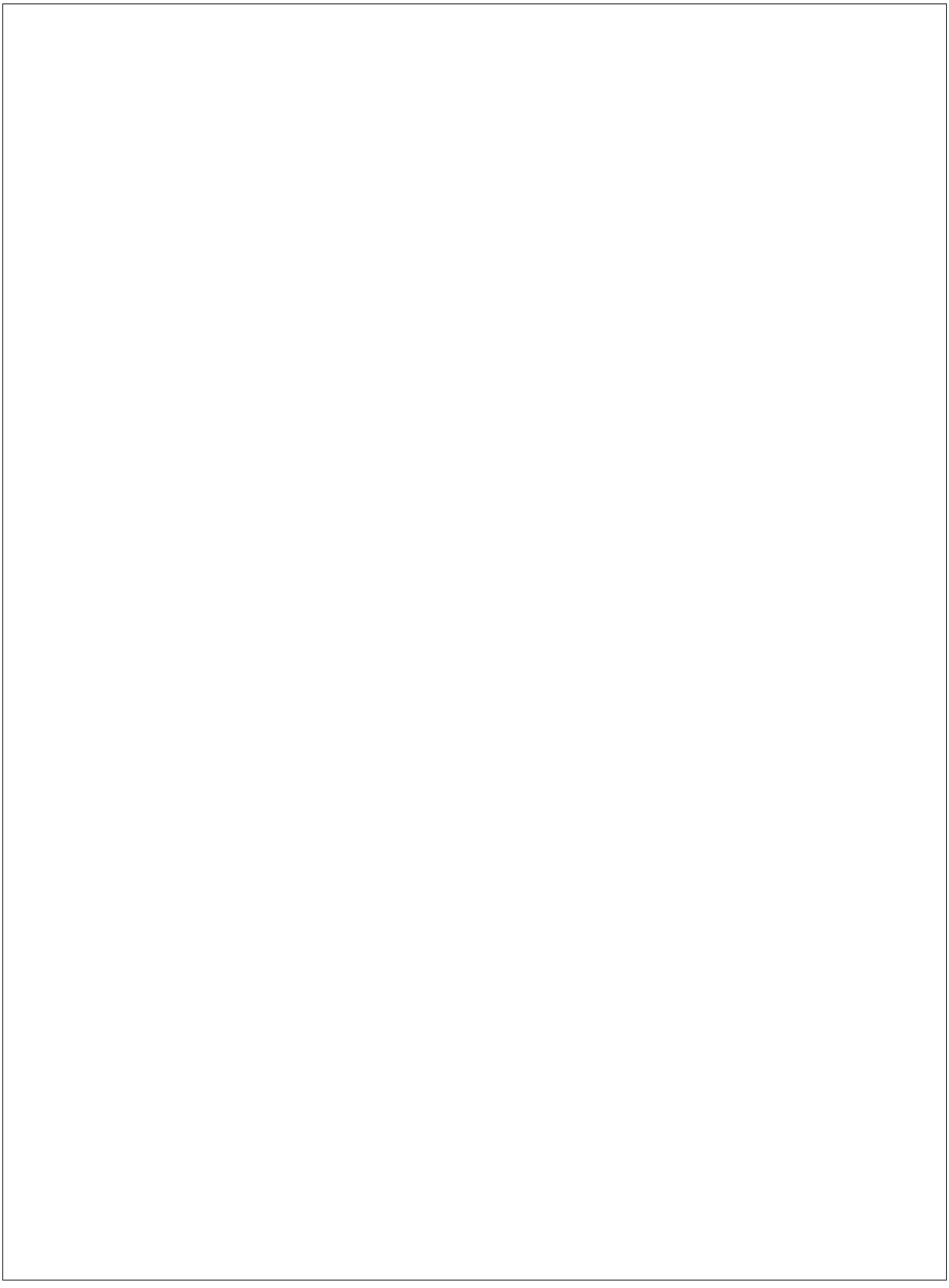
. . . /10

Question 10

TID	items bought
1	{dinde, bière, ail, coca}
2	{bière, endive, ail}
3	{dinde, bière, ail, coca}
4	{dinde, endive, ail, coca}
5	{bière, endive, ail, coca}
6	{bière, ail, coca}
7	{endive, ail}
8	{dinde, bière, endive}
9	{dinde, ail, coca}
10	{bière, ail}

Le support seuil est de 0.25 (c'est-à-dire, 25%). Montrez les résultats (structures de données et calculs) de l'exécution de l'algorithme FP-growth sur ce jeu de données. Il ne faut pas montrer l'exécution complète; arrêtez-vous au moment où **quatre** *frequent itemsets* ont été trouvés.

. . . /9



Donnez les *maximal frequent itemsets*.

. . . /3

Donnez les *nonclosed frequent itemsets*, i.e. les *frequent itemsets* qui ne sont pas *closed* (il y en a quatre). Expliquez pourquoi ces quatre *frequent itemsets* ne sont pas *closed*.

. . . /3

Question 11 Soit $S = \{(0,7), (3,14), (9,0), (14,6), (20,2)\}$ un ensemble de point en deux dimensions.

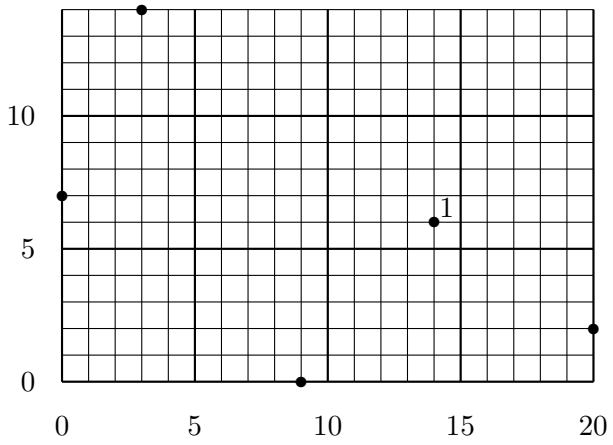
1. Exécutez l'algorithme de Dasgupta et Long, avec $\beta = 2$ et en prenant $(14,6)$ comme le point numéroté 1. Donnez les détails des calculs.

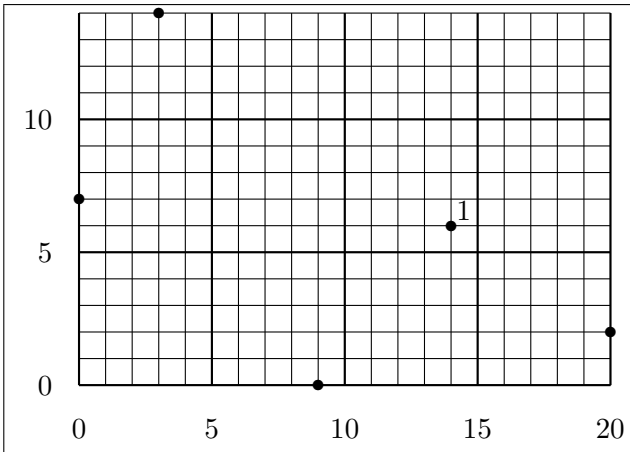
2. Comparez le coût du 3-clustering obtenu avec le coût du 3-clustering optimal, où le coût d'un clustering est le **diamètre maximal** de ses clusters. Est-ce que l'on reste bien "dans un facteur 8" ?

Voici la matrice des distances euclidiennes:

	$(0,7)$	$(3,14)$	$(9,0)$	$(14,6)$	$(20,2)$
$(0,7)$	0	$\sqrt{58} = 7.62$	$\sqrt{130} = 11.4$	$\sqrt{197} = 14.04$	$\sqrt{425} = 20.62$
$(3,14)$		0	$\sqrt{232} = 15.23$	$\sqrt{185} = 13.6$	$\sqrt{433} = 20.81$
$(9,0)$			0	$\sqrt{61} = 7.81$	$\sqrt{125} = 11.18$
$(14,6)$				0	$\sqrt{52} = 7.21$
$(20,2)$					0

.../15




10
5
0
1
0
5
10
15
20