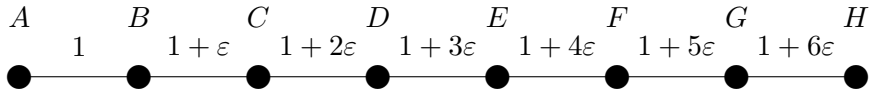
Data Mining et Data Warehousing, 4 juin 2010

Cahier fermé. Durée : 3 heures

Nom et prénom
Année

Question 1 A, B, ..., H sont huit points. La distance entre A et B est 1. La distance entre B et C est $1+\varepsilon$, avec ε une distance infinitésimale (i.e. minuscule). La distance entre C et D est $1+2\varepsilon$. La distance entre D et E est $1+3\varepsilon$. Etc.



Donnez le résultat des algorithmes *Single Link* et *Complete Link* pour ce jeu de données, en complétant le tableau suivant.

. . . /10

	Single Link	**Complete Link**
8-clustering	$\{A\},\{B\},\{C\},\{D\},\{E\},\{F\},\{G\},\{H\}$	$\{A\},\{B\},\{C\},\{D\},\{E\},\{F\},\{G\},\{H\}$
7-clustering		
6-clustering		
5-clustering		
4-clustering		
3-clustering		
2-clustering		
1-clustering	$\{A,B,C,D,E,F,G,H\}$	$\{A,B,C,D,E,F,G,H\}$

Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Question 2 Cophenetic distance.

.../5

Question 3 Ordinal attribute.

.../5

Question 4 Oblique decision tree.

. . . /5

Question 5 Vertical data layout of transaction data set.

. . . /5

Question 6 Soit $S = \{(0,0),(1,2),(1,6),(3,5),(3,7),(3,10)\}$ un ensemble de points en deux dimensions. Exécutez l'algorithme de Dasgupta et Long, en prenant $(1,2)$ comme le point numéroté 1. **Au cas où *farthest first traversal* laisse le choix entre plusieurs points, sélectionnez le point qui est le plus proche du point point numéroté 1.** Voici la matrice des distances euclidiennes:

	$(0,0)$	$(1,2)$	$(1,6)$	$(3,5)$	$(3,7)$	$(3,10)$
$(0,0)$	0	2.24	6.08	5.83	7.62	10.44
$(1,2)$	2.24	0	4	3.61	5.39	8.25
$(1,6)$	6.08	4	0	2.24	2.24	4.47
$(3,5)$	5.83	3.61	2.24	0	2	5
$(3,7)$	7.62	5.39	2.24	2	0	3
$(3,10)$	10.44	8.25	4.47	5	3	0

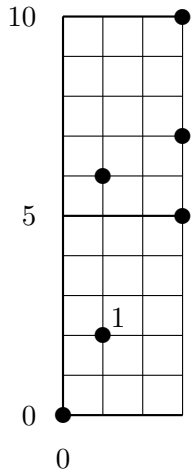
Donnez les détails des calculs. Puis complétez le tableau suivant, où le coût est le **diamètre maximal**. Est-ce que l'on s'approche du "facteur 8" ?

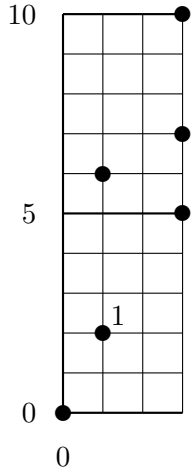
.../20

	Dasgupta et Long	Coût	**k-clustering optimal**	Coût
6-clustering	$\{(0,0)\},\{(1,2)\},\{(1,6)\},$ $\{(3,5)\},\{(3,7)\},\{(3,10)\}$	0	$\{(0,0)\},\{(1,2)\},\{(1,6)\},$ $\{(3,5)\},\{(3,7)\},\{(3,10)\}$	0
5-clustering				
4-clustering				
3-clustering			$\{(0,0),(1,2)\},\{(3,10)\}$ $\{(1,6),(3,5),(3,7)\}$	2.24
2-clustering				
1-clustering	$\{(0,0),(1,2),(1,6),$ $(3,5),(3,7),(3,10)\}$	10.44	$\{(0,0),(1,2),(1,6),$ $(3,5),(3,7),(3,10)\}$	10.44

Le facteur maximal observé est :

Détails des calculs.





Expliquez chaque figure de façon détaillée. Évitez des explications trop générales qui ne sont pas spécifiques pour la figure en question. Par exemple, pour la figure 4, il faut entrer dans les détails de la figure : que signifient les chiffres ? quel est le lien entre les deux arbres ? qu'est-ce que les auteurs ont voulu montrer avec cette image ?

Question 7 Figure 1.

. . . /10

Question 8 Figure 2.

. . . /10

Question 9 Figure 3.

. . . /10

Question 10 Figure 4.

. . . /10

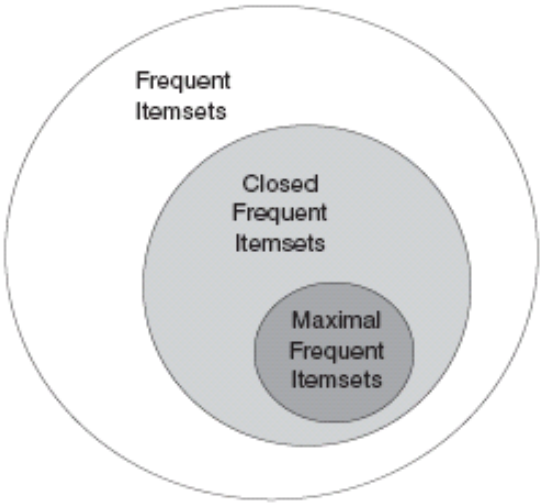


Figure 6.18. Relationships among frequent, maximal frequent, and closed frequent itemsets.

Figure 1

5. Decision trees provide an expressive representation for learning discrete-valued functions. However, they do not generalize well to certain types of Boolean problems. One notable example is the parity function, whose value is 0 (1) when there is an odd (even) number of Boolean attributes with the value $True$. Accurate modeling of such a function requires a full decision tree with 2^d nodes, where d is the number of Boolean attributes (see Exercise 1 on page 198).

Figure 2

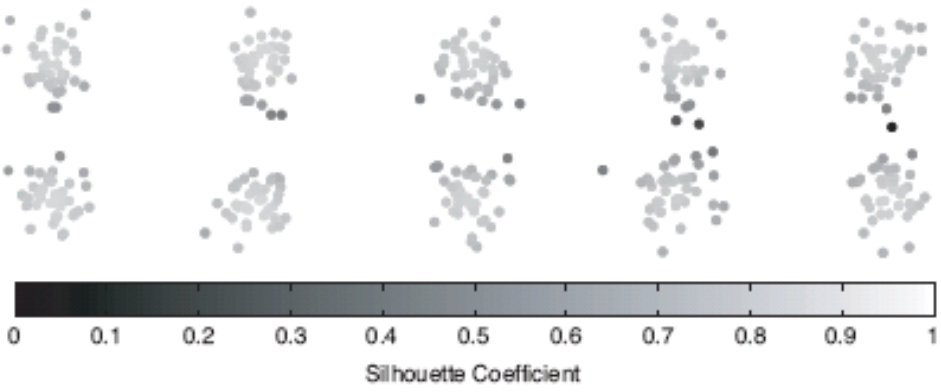


Figure 8.29. Silhouette coefficients for points in ten clusters.

Figure 3

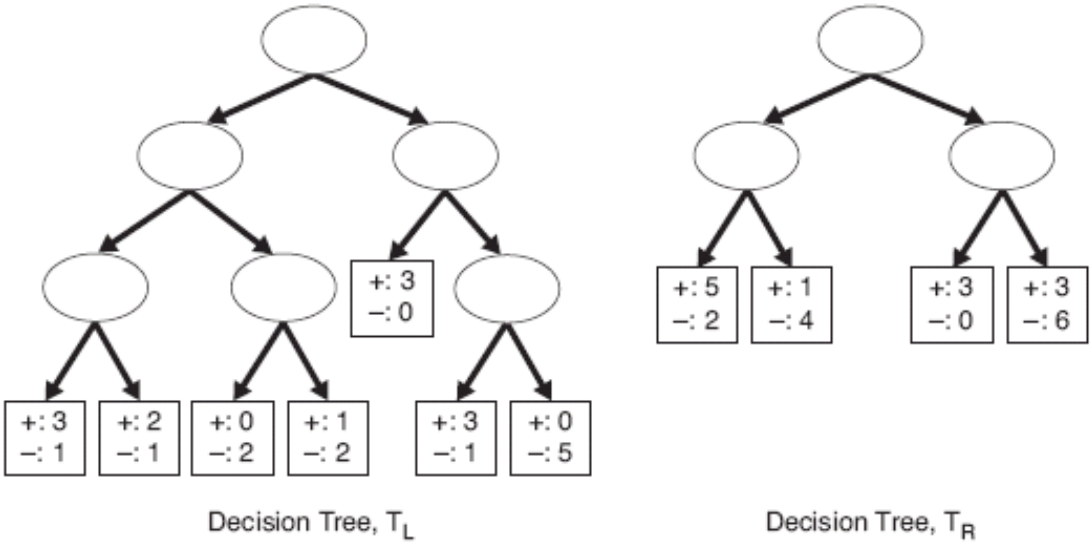


Figure 4.27. Example of two decision trees generated from the same training data.

Figure 4