USTHB
Faculté d'Electronique et Informatique
Département d'Informatique
Master S2I, 2ème année

30 Janvier 2014

Examen de rattrapage en Data mining

Exercice 1 (10 pts)

Le tableau suivant contient des données d'apprentissage d'une base de données d'employés. Les données ont été généralisées, par exemple, "31..35" pour l'âge représente la tranche d'âge de 31 à 35 ans. Pour une entrée de ligne donnée, *compte* représente le nombre de lignes de données qui ont pour valeurs *département*, *statut*, *âge* et *salaire* celles données dans la ligne correspondante. Soit *Etat* l'attribut de classe.

Département	Etat	Age	Salaire	compte
Ventes	Sénior	31..35	46K..50K	30
Ventes	Junior	26..30	26K..30K	40
Ventes	Junior	31..35	31K..35K	40
systèmes	Junior	21. .25	46K..50K	20
systèmes	Sénior	31. . .35	66K..70K	5
systèmes	Junior	26. . .30	46K..50K	3
systèmes	Sénior	41. . .45	66K..70K	3
marketing	Sénior	36. . .40	46K..50K	10
marketing	Junior	31. . .35	41K..45K	4
secrétariat	Sénior	46. . .50	36K..40K	4
secrétariat	Junior	26. . .30	26K..30K	6

1) Rappeler l'algorithme de construction de l'arbre de décision. (2 pts)
2) Comment l'algorithme C4.5 diffère-t-il de l'algorithme ID3 ? (2 pts)
3) Comment modifier ID3 pour prendre en considération le nombre de tuples identiques? (2 pts)
4) Utilisez votre algorithme pour construire un arbre de décision pour les données du tableau ci-dessus. (4 pts)

Exercice 2 (10 pts)

Considérer un plan géométrique à deux dimensions et les points de coordonnées suivants:
$A_1(1, 2)$, $A_2(4,6)$, $A_3(9, 2)$, $A_4(5,8)$, $A_5(2, 3)$, $A_6(8, 1)$, $A_7(5, 6)$, $A_8(10, 7)$, $A_9(2, 4)$, $A_{10}(5, 7)$.
On s'intéresse à appliquer l'algorithme k-means pour regrouper les points en 3 clusters.

1) définir une distance permettant l'application de l'algorithme. Justifier. (1 pt)
2) Appliquer l'algorithme k-means. Montrer clairement toutes les étapes. (4 pts)
3) Illustrer le résultat graphiquement. Retrouver les clusters. (1 pt)
4) Proposer trois points de démarrage de l'algorithme k-means, qui permettront d'optimiser le temps d'exécution. (2 pts)
5) Appliquer graphiquement l'algorithme k-means avec ces trois points. (2 pts)

Rédiger vos réponses clairement et de manière précise.