USTHB, Faculté d'Electronique et Informatique
Département d'Informatique
Master S2I, 2ème année
Module : Data Mining

TD 4

Exercice 1

1) Ecrire l'algorithme de génération de l'arbre de décision.
2) Calculer sa complexité.
3) Quelles sont les trois mesures les plus populaires utilisées dans l'algorithme de l'arbre de décision ?
4) Dresser un tableau comparatif de ces mesures. Quel est l'inconvénient majeur de l'algorithme de l'arbre de décision ?
5) Citer deux méthodes qui peuvent pallier à cet inconvénient.

Exercice 2

Le tableau suivant contient une base de données d'employés. Certaines données ont été regroupées dans des intervalles, par exemple, "31.. 35" pour l'âge représente la tranche d'âge de 31 à 35 ans. La colonne 'nombre' représente le nombre d'occurrence de l'instance.

département	*statut*	*âge*	*salaire*	*nombre*
ventes	senior	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	senior	31..35	66K..70K	5
systèmes	junior	26..30	46K..50K	3
systèmes	senior	41..45	66K..70K	3
marketing	senior	36..40	46K..50K	10
marketing	junior	31..35	41K..45K	4
secrétariat	senior	46..50	36K..40K	4
secrétariat	junior	26..30	26K..30K	6

En considérant l'attribut statut comme attribut de classe, engendrer l'arbre de décision de ces données sans tenir compte de la colonne 'nombre'.

1) Comment modifier l'algorithme de l'arbre de décision pour prendre en compte le nombre d'occurrence de l'instance ?
2) Déduire l'arbre de décision engendré de l'exécution de l'algorithme modifié.

Exercice 3.

Considérer la base d'apprentissage 'jouer au tennis' suivante :

Day	Outlook	Temperature	Humidity	Wind	*PlayTennis*?
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

Soit l'instance **New day = (sunny, cool, high, light)** à classer. Pour déterminer la classe de l'instance **New day :**

1) Appliquer la méthode de la classification Bayésienne naïve.
2) Proposer une mesure de similarité entre les instances.
3) Appliquer l'algorithme k-NN pour k=3.