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

Corrigé-Type du TD 4

Exercice1

1) Ecrire l'algorithme de génération de l'arbre de décision.

procédure Arbre-décision (D',C,AT')
Input : base d'apprentissage D', C attribut classe ; AT' ensemble des attributs ;
Output : racine de l'arbre de décision ADD ;
début
tant que (AT' ≠ {}) **et** (D' ≠ {}) **faire**
début
Pour chaque classe C_i **faire**
p_i := probabilité qu'une instance de D appartienne à la classe C_i ;
$Info(D') := -\sum_{i=1}^{i=m} p_i \log_2 p_i$; (*m est le nombre de classes*)
pour chaque attribut A **faire**
début
$Info_A(D') := \sum_{j=1}^{j=k} \frac{|D_j|}{|D'|} Info(D_j)$; (*k est le nombre de valeurs de A*)
$Gain(A) := Info(D') - Info_A(D')$;
fin ;
sélectionner A := rechercher(attribut, maximum gain) ;
créer-nœud(A) ;
pour chaque valeur possible v_i de A **faire**
début nœud := Créer-nœud(v_i) ;
Créer un lien de A vers nœud ;
Soit Exemples(v_i) := sous-ensemble de D' qui a la valeur v_i pour A ;
si Exemples(v_i) est vide **alors**
Valeur du nœud := valeur de l'attribut classe pour Exemples(v_i);
sinon valeur du nœud := Arbre-décision(Exemples(v_i), C, AT'-{A});
fin ;
retourner (A) ;
fin

programme principal ;
Input : base d'apprentissage D, C attribut classe ; AT ensemble des attributs ;
Output : racine de l'arbre de décision ADD ;
début
racine := Arbre-décision(D, C, AT);
fin ;

2) Calculer sa complexité.
L'arbre est construit niveau par niveau. Au pire cas, chaque niveau sera représenté par un attribut. Dans ce cas le nombre maximum de nœuds sera égal $\prod_1^{\#attributs} |v(A_i)|$. La complexité est donc $O(\prod_1^{\#attributs} |v(A_i)|)$.

3) Quelles sont les trois mesures les plus populaires utilisées dans l'algorithme de l'arbre de décision ?

Information Gain

Gain ratio

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

GainRatio(A) = Gain(A)/SplitInfo(A)

Gini index

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

4) Dresser un tableau comparatif de ces mesures.

	Avantage	Inconvénient
Gain information	Relativement rapide Simple à interpréter	couteux pour les attributs à plusieurs valeurs
Gain ratio	Plus fiable pour les attributs avec peu de valeurs	a tendance à préférer les divisions non équilibrées dans lesquelles une partition est beaucoup plus petite que les autres
Gini index		1. couteux pour les attributs à plusieurs valeurs 2. Couteux en temps : Doit énumérer tous les points de coupure pour chaque attribut

Quel est l'inconvénient majeur de l'algorithme de l'arbre de décision ?
L'inconvénient majeur est que l'algorithme aura des difficultés à s'exécuter dans le cas des données massives et plus particulièrement dans le cas multidimensionnel et multi-valeurs.

5) Citer deux méthodes qui pallient à cet inconvénient.
Pour pallier à ce problème, il faut penser aux méthodes suivantes :
Rainforest
BOAT

Exercice 2

Le tableau suivant contient une base de données d'employés. Certaines données ont été groupées 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'exemples de données ayant les valeurs pour département, statut, âge et salaire indiquées dans la ligne.

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

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

Class P : statut = senior

Class N : statut = junior

$$Info(D) := -\frac{5}{11}\log_2\frac{5}{11}-\frac{6}{11}\log_2\frac{6}{11} = (-0.45)*(-1.16) - 0.54*(-0.89) = 1$$

âge	senior	junior	I(senior, junior)
31..35	2	2	-1
26..30	0	3	0
21..25	0	1	0
41..45	1	0	0
36..40	1	0	0
46..50	1	0	0

$$Info_{age}(D) = \frac{4}{11}I(2,2) + \frac{3}{11}I(0,3) + \frac{1}{11}I(0,1) + \frac{1}{11}I(1,0) + \frac{1}{11}I(1,0) + \frac{1}{11}I(1,0)$$

$$I(2,2) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = -\log_2\frac{1}{2} = 1$$

$$I(0,3) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$$

$$Info_{age}(D) = \frac{4}{11}1 + \frac{3}{11}0 + \frac{1}{11}0 + \frac{1}{11}0 + \frac{1}{11}0 + \frac{1}{11}0 = 0.36$$

$$Gain(\hat{a}ge) = 1 - 0.36 = 0.64$$

salaire	senior	junior	I(senior, junior)
46K..50K	2	2	-1
26K..30K	0	2	0
31K..35K	0	1	0
66K..70K	2	0	0
41K..45K	0	1	0
36K..40K	1	0	0

$$Info_{salaire}(D) = \frac{4}{11}1 + \frac{2}{11}0 + \frac{1}{11}0 + \frac{2}{11}0 + \frac{1}{11}0 + \frac{1}{11}0 = 0.36$$

$$Gain(salaire) = 1 - 0.36 = 0.64$$

département	senior	junior	I(senior, junior)
ventes	1	2	0.92
systèmes	2	2	1
marketing	1	1	1
secrétariat	1	1	1

$$I(1,2) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.33x1.6 + 0.66x\ 0.6 = 0.92$$

$$I(2,2) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

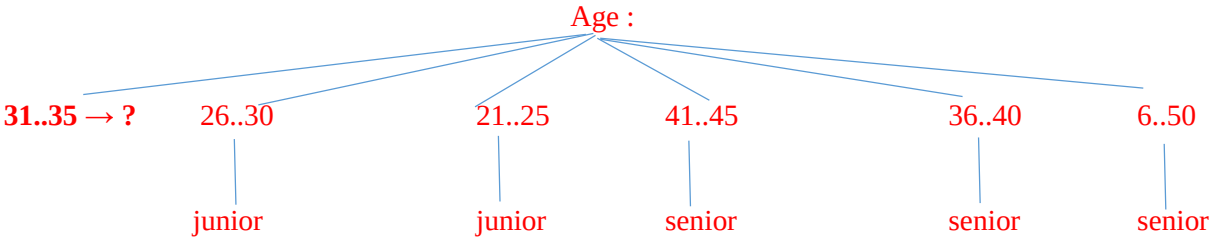
$$I(1,1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Info_{département}(D) = -\frac{3}{11}I(1,2) + \frac{4}{11}I(2,2) + \frac{2}{11}I(1,1) + \frac{2}{11}I(1,1)$$

$$= \frac{3}{11}0.92 + \frac{4}{11} + \frac{2}{11} + \frac{2}{11} = 0.27x\ 0.92 + 0.72 = 0.96$$

$$Gain(département) = 1 - 0.96 = 0.04$$

0n sélectionne âge ou salaire :



Pour la tranche d'âge 31..35, il reste les exemples suivants :

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

$$Info(D) := -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Info_{département}(D) = \frac{2}{4}I(1,1) + \frac{1}{4}I(1,0) + \frac{1}{4}I(0,1) = \mathbf{0.5}$$

$$Gain(département) = 1 - 0.5 = 0.5$$

$$Info_{salaire}(D) = \frac{1}{4}I(1,0) + \frac{1}{4}I(0,1) + \frac{1}{4}I(1,0) + \frac{1}{4}I(0,1) = 0$$

$$Gain(salaire) = 1 - 0 = \mathbf{1}$$

On choisit alors salaire et on obtient :

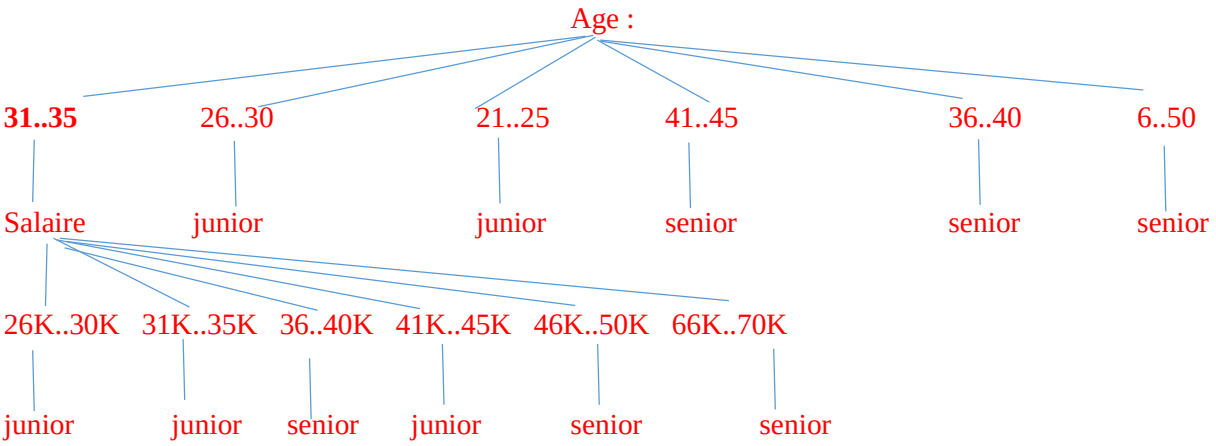
46K..50K → senior

31K..35K → junior

66K..70K→ senior

41K..45K → junior

On obtient l'arbre suivant:



2) Comment modifier l'algorithme de l'arbre de décision pour prendre en compte le nombre d'instances ayant les mêmes valeurs des attributs que celles indiquées dans la ligne ?

Le nombre total des exemples sera égal à la somme des nombres d'instances de chaque ligne. Le calcul de l'heuristique 'gain information' sera :

$$Info(D) := -\sum_{i=1}^{i=m} p_i \log_2 p_i \ ;$$

Où $p_i = \frac{nb\ d'exemples\ appartenant\ à\ la\ classe\ Ci}{\#exemples}$

$$Info_A(D) := \sum_{j=1}^{j=k} \frac{|D_j|\, n_j}{|D|} x Info(D_j\, n_j) \ ;$$

$$Gain(A) := Info(D) - Info_A(D) \ ;$$

$|D| = \sum_1^{\#exemples} n_i$

$$\#exemples = \sum_{i=1}^{i=taille(table)} nombre(i)$$

3) Déduire l'arbre de décision de l'exécution de l'algorithme modifié.

$Info(D) := -\frac{52}{165}\log_2\frac{52}{165}-\frac{113}{165}\log_2\frac{113}{165}$=(-0.31)(-1.69)-(-0.68)(-0.56)=0.90

Pour notre donnée, on aura 30+40+40+20+5+3+3+10+4+4+6=165.

âge	senior	junior	I(senior, junior)
31..35	35	44	0.98
26..30	0	49	0
21..25	0	20	0
41..45	3	0	0
36..40	10	0	0
46..50	4	0	0

$$Info_{age}(D) = \frac{79}{165}I(35,44) + \frac{49}{165}I(0,49) + \frac{20}{165}I(0,20) + \frac{3}{165}I(3,0) + \frac{10}{165}I(10,0) + \frac{4}{165}I(4,0)$$

$$I(35,44) = -\frac{35}{79}\log_2\frac{35}{79} - \frac{44}{79}\log_2\frac{44}{79} = 0.44x1.18 + 0.55x0.83 = 0.51 + 0.47 = 0.98$$

$$Info_{age}(D) = \frac{79}{165}0.98 + \frac{49}{165}0 + \frac{20}{165}0 + \frac{3}{165}0 + \frac{10}{165}0 + \frac{4}{165}0 = 0.47$$

$$Gain(\hat{a}ge) = info(D) - Info_{age}(D) = 0.90 - 0.47 = \mathbf{0.43}$$

salaire	senior	junior	I(senior, junior)
46K..50K	40	23	0.94
26K..30K	0	46	0
31K..35K	0	40	0
66K..70K	8	0	0
41K..45K	0	4	0
36K..40K	4	0	0

$$I(40,23) = -\frac{40}{63}\log_2\frac{40}{63} - \frac{23}{63}\log_2\frac{23}{63} = 0.63x0.66 + 0.36x1.47 = 0.94$$

$$Info_{salaire}(D) = \frac{63}{165}0.94 + \frac{46}{165}0 + \frac{40}{165}0 + \frac{8}{165}0 + \frac{4}{165}0 + \frac{4}{165}0 = 0.35$$

$$Gain(salaire) = info(D) - Info_{salaire}(D) = 0.90 - 0.35 = \mathbf{0.55}$$

département	senior	junior	I(senior, junior)
ventes	30	80	0.85
systèmes	8	23	0.83
marketing	10	4	0.86
secrétariat	4	6	0.98

$$I(30,80) = -\frac{30}{110}\log_2\frac{30}{110} - \frac{80}{110}\log_2\frac{80}{110} = 0.27x1.89 + 0.72x0.47 = 0.51 + 0.34 = 0.85$$

$$I(8,23) = -\frac{8}{31}\log_2\frac{8}{31} - \frac{23}{31}\log_2\frac{23}{31} = 0.26x1.95 + 0.74x0.43 = 0.83$$

$$I(10,4) = -\frac{10}{14}\log_2\frac{10}{14} - \frac{4}{14}\log_2\frac{4}{14} = 0.71x0.49 + 0.28x1.84 = 0.86$$

$$I(4,6) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} = 0.4x1.33 + 0.6x0.74 = 0.98$$

$$Info_{département}(D) = \frac{110}{165}0.85 + \frac{31}{165}0.83 + \frac{14}{165}0.86 + \frac{10}{165}0.98$$
$$= 0.66x0.85 + 0.19x0.83 + 0.08x0.86 + 0.06x0.98 = 0.56 + 0.15 + 0.07 + 0.06 = 0.84$$

$$Gain(département) = info(D) - Info_{département}(D) = 0.90 - 0.84 = \mathbf{0.06}$$

0n sélectionne alors ***salaire***

46K..50K → ?
26K..30K → junior
31K..35K → junior
66K..70K → senior
41K..45K → junior
36K..40K → senior
46K..50K → ?

département	*statut*	*âge*	*salaire*	*nombre*
ventes	senior	31..35	46K..50K	30
systèmes	junior	21..25	46K..50K	20
systèmes	junior	26..30	46K..50K	3
marketing	senior	36..40	46K..50K	10

$$Info(D) := -\frac{40}{63}\log_2\frac{40}{63} - \frac{23}{63}\log_2\frac{23}{63} = 0.63x0.66 + 0.36x1.47 = 0.94$$

Age

âge	**senior**	**junior**	**I(senior, junior)**
31..35	30	0	0
26..30	0	3	0
21..25	0	20	0
36..40	10	0	0

$$Info_{age}(D) = 0$$

$$Gain(âge) = info(D) - Info_{age}(D) = 0.94 - 0 = \mathbf{0.94}$$

Département

département	**senior**	**junior**	**I(senior, junior)**
ventes	30	0	0
systèmes	0	23	0
marketing	10	0	0

$$Gain(département) = info(D) - Info_{age}(D) = 0.94 - 0 = \mathbf{0.94}$$

On sélectionne soit âge soit département

Département

ventes → senior
systèmes → junior
marketing → senior

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.

P(Ci) : P(PlayTennis=No) = 5/14=0.357
P(PlayTennis=Yes) = 9/14 = 0.643

P(**New day** /Ci) pour chaque classe :
P(Outlook = sunny | PlayTennis=No)=3/5=0.6
P(Outlook = sunny | PlayTennis=Yes)=2/9=0.222
P(Temperature = cool | PlayTennis=No)=1/5=0.2
P(Temperature = cool | PlayTennis=Yes)=3/9=0.333
P(Humidity = high | PlayTennis=No)=4/5=0.8
P(Humidity = high | PlayTennis=Yes)=3/9=0.333
P(Wind = light | PlayTennis=No)=2/5=0.4
P(Wind = light | PlayTennis=Yes)=6/9=0.666

P(**New day** / PlayTennis=No) = 0.6 * 0.2 * 0.8 * 0.4 =0.0384
P(**New day** / PlayTennis=Yes) = 0.222 * 0.333 * 0.333 * 0.666 =0.0163

P(**New day** / PlayTennis=No)* P(PlayTennis=No) = 0.0384 * 0.357 = 0.0137
P(**New day** / PlayTennis=No)* P(PlayTennis=Yes) = 0.0163 * 0.643 = 0.0104

New day appartient **à** PlayTennis=No

2) Proposer une mesure de similarité entre les instances.

Similarity(i,j) = nombre de valeurs d'attributs identiques entre i et j / nombre d'attributs

3) Appliquer l'algorithme k-NN pour k=3.

New day = (sunny, cool, high, light)

Distance(New day, 1) = similarity(New day, 1) = 3/4= 0.75
Distance(New day, 2) = 2/4= 0.5
Distance(New day, 3) = 2/4 = 0.5
Distance(New day, 4) = 2/4 = 0.5
Distance(New day, 5) = 2/4 = 0.5
Distance(New day, 6) = 1/4 = 0.25
Distance(New day, 7) = 1/4 = 0.25
Distance(New day, 8) = 3/4 = 0.75
Distance(New day, 9) = 3/4 = 0.75
Distance(New day, 10) = 1/4 = 0.25
Distance(New day, 11) = 1/4 = 0.25
Distance(New day, 12) = 1/4 = 0.25
Distance(New day, 13) = 1/4 = 0.25
Distance(New day, 14) = 1/4 = 0.25

Les 3 plus proches voisins de New day = (1, PlayTennis=No), (8, PlayTennis=No), (9, PlayTennis=Yes)

New day appartient donc **à** PlayTennis=No