Université de Chlef
Département Informatique
Filière : 2ème Année Master - ISIA et IL

Janvier 2018

Examen semestriel

Modules "**Fouille et extraction de données" & "Datamining"**

Durée : 01H30

Corrigé

Exercice 1 (10 points) :

Soit l'ensemble D des entiers suivants :

D= { 2, 5, 8, 10, 11, 18, 20 }

On veut répartir les données de D en trois (3) clusters, en utilisant l'algorithme Kmeans. La distance d entre deux nombres a et b est calculée ainsi :

d(a , b) = |a - b| (la valeur absolue de a moins b)

Travail à faire :

1/ Appliquez Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul.

Réponse :

Initialisation :

des centres de gravité :	$\mu1=8$	$\mu2=10$	$\mu3=11$
des clusters :	$C1=\emptyset$	$C2=\emptyset$	$C3=\emptyset$

Itération 1 :

Calcul des distances

Nombre 2 :

$d(2, \mu1)=|2-8|=\mathbf{6}$

$d(2, \mu2)=|2-10|=8$

$d(2, \mu3)=|2-11|=9$

2 est affecté au cluster C1.

Nombre 5 :

$d(5, \mu1)=|5-8|=\mathbf{3}$

$d(5, \mu2)=|5-10|=5$

$d(5, \mu3)=|5-11|=6$

5 est affecté au cluster C1.

Nombre 8 :

$d(8, \mu1)=|8-8|=\mathbf{0}$

$d(8, \mu2)=|8-10|=2$

$d(8, \mu3)=|8-11|=3$

8 est affecté au cluster C1.

Nombre 10 :

d(10, μ1)=|10-8|=2

*d(10, μ2)=|10-10|=**0***

d(10, μ3)=|10-11|=1

10 est affecté au cluster C2.

Nombre 11 :

d(11, μ1)=|11-8|=3

d(11, μ2)=|11-10|=1

*d(11, μ3)=|11-11|=**0***

11 est affecté au cluster C3.

Nombre 18 :

d(18, μ1)=|18-8|=10

d(18, μ2)=|18-10|=8

*d(18, μ3)=|18-11|=**7***

18 est affecté au cluster C3.

Nombre 20 :

d(20, μ1)=|20-8|=12

d(20, μ2)=|20-10|=10

*d(20, μ3)=|20-11|=**9***

20 est affecté au cluster C3.

Mise à jour des clusters :

C1={ 2, 5, 8} *C2={10}* *C3={11, 18, 20}*

R- estimation des centres de gravité :

μ1= (2+5+8)/3 *μ2=10/1* *μ3=(11+18+20)/3*

μ1=5 *μ2=10* *μ3=16.33*

(2 points)

Itération 2 :

Calcul des distances

Nombre 2 :

*d(2, μ1)=|2-5|=**3***

d(2, μ2)=|2-10|=8

d(2, μ3)=|2-16.33|=14.33

2 est affecté au cluster C1.

Nombre 5 :

*d(5, μ1)=|5-5|=**0***

d(5, μ2)=|5-10|=5

d(5, μ3)=|5-16.33|=11.33

5 est affecté au cluster C1.

Nombre 8 :

d(8, μ1)=|8-5|=3

*d(8, μ2)=|8-10|=**2***

d(8, μ3)=|8-16.33|=8.33

8 est affecté au cluster C2.

Nombre 10 :

d(10, μ1)=|10-5|=5

*d(10, μ2)=|10-10|=**0***

d(10, μ3)=|10-16.33|=6.33

10 est affecté au cluster C2.

Nombre 11 :

d(11, μ1)=|11-5|=6

*d(11, μ2)=|11-10|=**1***

d(11, μ3)=|11-16.33|=5.33

11 est affecté au cluster C2.

Nombre 18 :

d(18, μ1)=|18-5|=13

d(18, μ2)=|18-10|=8

*d(18, μ3)=|18-16.33|=**1.67***

18 est affecté au cluster C3.

Nombre 20 :

d(20, μ1)=|20-5|=15

d(20, μ2)=|20-10|=10

*d(20, μ3)=|20-16.33|=**3.67***

20 est affecté au cluster C3.

Mise à jour des clusters :

C1={ 2, 5} *C2={8, 10, 11}* *C3={18, 20}*

R- estimation des centres de gravité :

μ1= (2+5)/2 *μ2=(8+10+11)/3* *μ3=(18+20)/2*

μ1=3.5 *μ2=9.66* *μ3=19*

(2 points)

Itération 3 :

Calcul des distances

Nombre 2 :

*d(2, μ1)=|2-3.5|=**1.5***

d(2, μ2)=|2-9.66|=7.66

d(2, μ3)=|2-19|=17

2 est affecté au cluster C1.

Nombre 5 :

*d(5, μ1)=|5-3.5|=**1.5***

d(5, μ2)=|5-9.66|=4.66

d(5, μ3)=|5-19|=14

5 est affecté au cluster C1.

Nombre 8 :

d(8, μ1)=|8-3.5|=4.5

*d(8, μ2)=|8-9.66|=**1.66***

d(8, μ3)=|8-19|=11

8 est affecté au cluster C2.

Nombre 10 :

d(10, μ1)=|10-3.5|=6.5

*d(10, μ2)=|10-9.66|=**0.34***

d(10, μ3)=|10-19|=9

10 est affecté au cluster C2.

Nombre 11 :

d(11, μ1)=|11-3.5|=7.5

*d(11, μ2)=|11-9.66|=**1.34***

d(11, μ3)=|11-19|=8

11 est affecté au cluster C2.

Nombre 18 :

d(18, μ1)=|18-3.5|=14.5

d(18, μ2)=|18-9.66|=8.34

*d(18, μ3)=|18-19|=**1***

18 est affecté au cluster C3.

Nombre 20 :

d(20, μ1)=|20-3.5|=16.5

d(20, μ2)=|20-9.66|=10.34

*d(20, μ3)=|20-19|=**1***

20 est affecté au cluster C3.

Mise à jour des clusters :

C1={ 2, 5} *C2={8, 10, 11}* *C3={18, 20}*

R- estimation des centres de gravité :

μ1= (2+5)/2 *μ2=(8+10+11)/3* *μ3=(18+20)/2*

μ1=3.5 *μ2=9.66* *μ3=19*

Stabilité : Les centres de gravité n'ont pas changé. L'algorithme s'arrête

(2 points)

2/ Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.

Réponse :

Les clusters résultats :

C1={ 2, 5} *C2={8, 10, 11}* *C3={18, 20}*

Nombre d'itérations = 3

(2 points)

3/ Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez.

Réponse :

Dans ce problème, les données sont ordonnées et restreintes dans un intervalle (de 2 à 20). Comme on veut construire 3 clusters, on est sûr que la borne inférieure (2) sera dans le cluster 1, et la borne supérieure (20) sera dans le cluster 3. Il est donc intéressant de choisir comme centres de gravité initiaux : la borne inférieure (2) pour le cluster 1, la borne supérieure (20) pour le cluster 3, et le milieu de l'intervalle (9) comme centre pour le cluster 2. Avec une telle initialisation, l'algorithme convergera après seulement 2 itérations.

(2 points)

Exercice 2 (10 points) :

Le tableau suivant contient des données sur les résultats obtenus par des étudiants de Tronc Commun (première année à l'Université). Chaque étudiant est décrit par 3 attributs : Est-il doublant ou non, la série du Baccalauréat obtenu et la mention. Les étudiants sont répartis en deux classes : Admis et Non Admis.

On veut construire un arbre de décision à partir des données du tableau, pour rendre compte des éléments qui influent sur les résultats des étudiants en Tronc Commun. Les lignes de 1 à 12 sont utilisées comme données d'apprentissage. Les lignes restantes (de 13 à 16) sont utilisées comme données de tests.

	Doublant	Série	Mention	Classe
1	Non	Maths	ABien	Admis
2	Non	Techniques	ABien	Admis
3	Oui	Sciences	ABien	Non Admis
4	Oui	Sciences	Bien	Admis
5	Non	Maths	Bien	Admis

6	Non	Techniques	Bien	Admis
7	Oui	Sciences	Passable	Non Admis
8	Oui	Maths	Passable	Non Admis
9	Oui	Techniques	Passable	Non Admis
10	Oui	Maths	TBien	Admis
11	Oui	Techniques	TBien	Admis
12	Non	Sciences	TBien	Admis
13	Oui	Maths	Bien	Admis
14	Non	Sciences	ABien	Non Admis
15	Non	Maths	TBien	Admis
16	Non	Maths	Passable	Non Admis

Travail à faire :

1/ Utiliser les données des lignes de 1 à 12 pour construire l'arbre en utilisant l'algorithme ID3. Montrez toutes les étapes de calcul. Dessinez l'arbre final.

Réponse :

On remarque que sur les 12 lignes des données d'apprentissage, 8 correspondent à la classe "Admis" et 4 à la classe "Non admis". L'entropie de l'ensemble S (à la racine de l'arbre) est donc égale à :

*Entropie (S)= - (8/12) * Log_2 (8/12) - (4/12) * Log_2 (4/12)*

Entropie (S)= 0.92

(0.5 point)

Pour connaitre quel attribut on doit choisir comme test au niveau de la racine de l'arbre, il faut calculer le gain d'entropie sur chacun des attributs : "Doublant", "Série" et "Mention".

Calcul du gain d'entropie sur l'attribut "Doublant" :

*Gain(S, Doublant) = Entropie(S) - 7/12 * Entropie(S_{Oui}) - 5/12 * Entropie(S_{Non})*

*avec Entropie(S_{Oui})= -3/7*Log_2(3/7) - 4/7*Log_2(4/7)*

*et Entropie(S_{Non})= -5/5*log_2(5/5)*

Gain(S, Doublant) =0.34

(0.5 point)

Calcul du gain d'entropie sur l'attribut "Série" :

*Gain(S, Série) = Entropie(S) - 4/12 * Entropie(S_{Maths}) - 4/12 * Entropie($S_{Techniques}$) - 4/12 * Entropie($S_{Sciences}$)*

Gain(S, Série) = 0.04

(0.5 point)

Calcul du gain d'entropie sur l'attribut "Mention" :

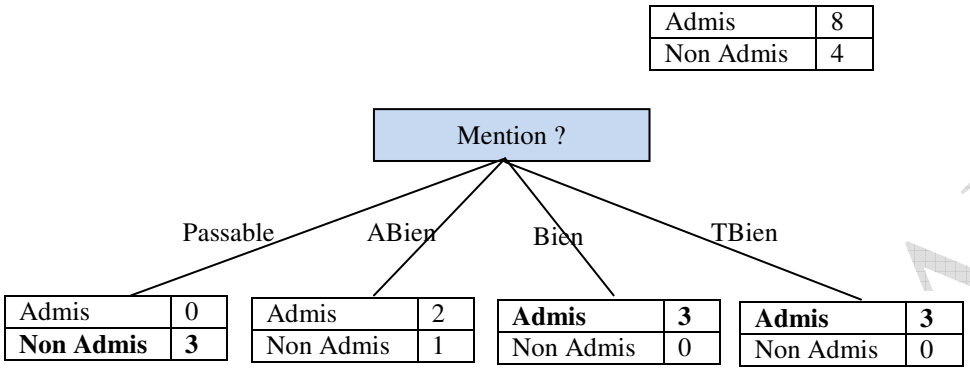
*Gain(S, Série) = Entropie(S) - 3/12 * Entropie($S_{Passable}$) - 3/12 * Entropie(S_{ABien}) - 3/12 * Entropie(S_{Bien}) - 3/12 * Entropie(S_{TBien})*

*Gain(S, Mention) = **0.69***

(0.5 point)

On constate que le plus grand gain d'entropie est obtenu sur l'attribut "Mention". C'est donc cet attribut qui est choisi comme test à la racine de l'arbre. Nous obtenons l'arbre partiel suivant :

(1 point)



(1 point)

On voit que mettre l'attribut "Mention" à la racine de l'arbre permet d'obtenir 4 branches dont 3 produisent des noeuds purs (finaux). Il ne reste à traiter que le nœud présentant un mélange correspondant à la branche "ABien". Ce nœud comporte un ensemble (que nous noterons S2) ayant 2 individus appartenant à la classe "Admis" et 1 individu de la classe "Non Admis". L'entropie de l'ensemble S2 est donc égale à :

*Entropie (S2)= - (2/3) * Log_2 (2/3) - (1/3) * Log_2 (1/3)*

Entropie (S2)= 0.92

(0.5 point)

Pour connaitre quel attribut on doit choisir comme test au niveau du nœud impur, il faut calculer le gain d'entropie sur chacun des attributs restants : "Doublant" et "Série".

Calcul du gain d'entropie sur l'attribut "Doublant" :

*Gain(S2, Doublant) = Entropie(S2) - 1/3 * Entropie(S_{Oui}) -2/3 * Entropie(S_{Non})*

*Gain(S2, Doublant) =**0.92***

(0.5 point)

Calcul du gain d'entropie sur l'attribut "Série" :

*Gain(S2, Série) = Entropie(S2) - 1/3 * Entropie(S_{Maths}) - 1/3 * Entropie($S_{Techniques}$) - 1/3 * Entropie($S_{Sciences}$)*

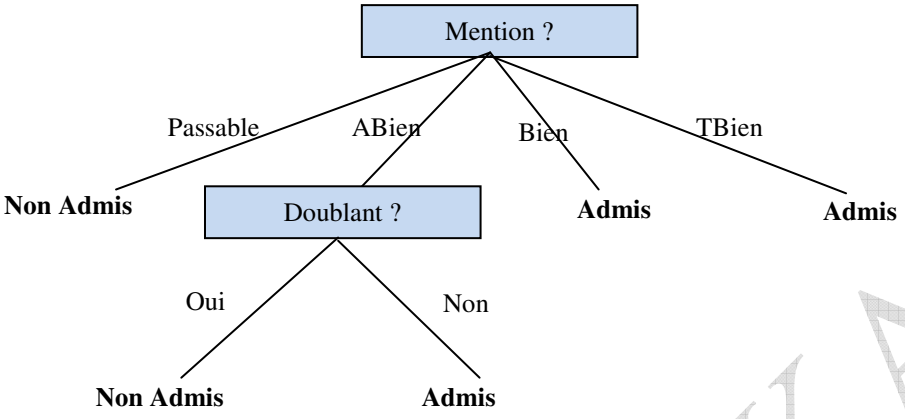
*Gain(S2, Série) = **0.92***

(0.5 point)

On constate que les deux attributs "Doublant" et "Série" procurent le même gain d'entropie. Nous pouvons donc choisir l'un ou l'autre comme test au niveau du nœud courant. Nous avons donc deux arbres de décision possibles :

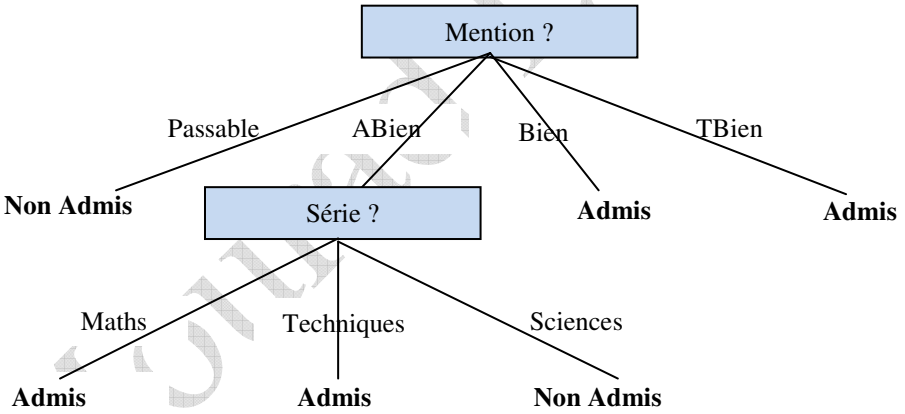
(1 point)

Premier arbre :



(1 point)

Deuxième arbre :



(1 point)

2/ Quels sont les résultats de test de l'arbre obtenu sur les données des lignes de 13 à 16 ?.

					Test de l'arbre 1		*Test de l'arbre 2*	
	Doublant	*Série*	*Mention*	*Classe*	*Classe déduite de l'arbre 1*	*Observation*	*Classe déduite de l'arbre 2*	*Observation*
13	*Oui*	*Maths*	*Bien*	*Admis*	*Admis*	*Correct*	*Admis*	*Correct*
14	*Non*	*Sciences*	*ABien*	*Non Admis*	*Admis*	*Erreur*	*Non Admis*	*Correct*
15	*Non*	*Maths*	*TBien*	*Admis*	*Admis*	*Correct*	*Admis*	*Correct*
16	*Non*	*Maths*	*Passable*	*Non Admis*	*Non Admis*	*Correct*	*Non Admis*	*Correct*

On remarque que l'arbre 1 a donné un taux d'erreur de 1/4 soit 25%, alors que l'arbre 2 présente un taux de succès de 100%. Cela suggère de retenir en définitif l'arbre 2 qui conforte l'idée suivante :

Les résultats obtenus par les étudiants de tronc commun sont déterminés par deux éléments : la mention obtenue de leur baccalauréat et la série. Les étudiants ayant une bonne mention (ABien ou plus) ou issus des filières Maths et Techniques ne trouvent pas de difficultés à passer la première année à l'Université.

(1.5 points)