Master Informatique

Fouille de données

Exercices : Arbres de décision

Exercice 1

Donner les arbres de décisions qui expriment les fonctions booléennes suivantes :

1. $A \wedge \neg B$
2. $A \vee (B \wedge C)$
3. $AXORB$
4. $(A \wedge B) \vee (C \wedge D)$

Exercice 2

jour	ciel	temp.	humidité	vent	jouer
1	soleil	chaud	élevée	faible	non
2	soleil	chaud	élevée	fort	non
3	couvert	chaud	élevée	faible	oui
4	pluie	doux	élevée	faible	oui
5	pluie	froid	normale	faible	oui
6	pluie	froid	normale	fort	non
7	couvert	froid	normale	fort	oui
8	soleil	doux	élevée	faible	non
9	soleil	froid	normale	faible	oui
10	pluie	doux	normale	faible	oui
11	soleil	doux	normale	fort	oui
12	couvert	doux	élevée	fort	oui
13	couvert	chaud	normale	faible	oui
14	pluie	doux	élevée	fort	non

Jouer au tennis

Construire l'arbre de décision en utilisant la fonction gain basée sur l'entropie.

Exercice 3

Une banque dispose des informations suivantes sur un ensemble de clients:

client	*M*	*A*	*R*	*E*	*I*
1	*moyen*	*moyen*	*village*	*oui*	*oui*
2	*élevé*	*moyen*	*bourg*	*non*	*non*
3	*faible*	*âgé*	*bourg*	*non*	*non*
4	*faible*	*moyen*	*bourg*	*oui*	*oui*
5	*moyen*	*jeune*	*ville*	*oui*	*oui*
6	*élevé*	*âgé*	*ville*	*oui*	*non*
7	*moyen*	*âgé*	*ville*	*oui*	*non*
8	*faible*	*moyen*	*village*	*non*	*non*

L'attribut client indique le numéro du client ; l'attribut M indique la moyenne des crédits sur le compte du client ; l'attribut A donne la tranche d'âge ; l'attribut R décrit la localité du client ; l'attribut E possède la valeur oui si le client possède un niveau d'études supérieur au bac ; l'attribut I (la classe) indique si le client effectue ses opérations de gestion de compte via Internet.

1) Quelle est l'entropie de la population ?
2) Pour la construction de l'arbre de décision, utilisez-vous l'attribut numéro de client ? Pourquoi?
3) Lors de la construction de l'arbre de décision, quel est l'attribut à tester à la racine de l'arbre ?
4) Construire l'arbre de décision complet.
5) Quel est le taux d'erreur de cet arbre estimé sur l'ensemble de clients 1 à 8 ?
6) Donner un intervalle de valeurs pour l'erreur réelle en utilisant une confiance de 90 %.

On se donne les 4 clients suivants :

test T	*M*	*A*	*R*	*E*	*I*
9	*moyen*	*âgé*	*village*	*oui*	*oui*
10	*élevé*	*jeune*	*ville*	*non*	*oui*
11	*faible*	*âgé*	*village*	*non*	*non*
12	*moyen*	*moyen*	*bourg*	*oui*	*non*

7) Comment chacun de ces clients est-il classé avec l'arbre de décision que vous avez proposé dans la question 4 ?
   Pour ces 4 clients, on apprend par ailleurs que les clients 9 et 10 gèrent leur compte par Internet, et que les clients 11 et 12 ne le font pas.
8) Quel est le taux d'erreur estimé sur les clients 9, 10, 11 et 12 ? Combien y a-t-il de faux positifs et de faux négatifs ?