

Série N°2 : Classification supervisée : Arbre de décision

Exercice N°1 :

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

1. $F(A, B) = A \wedge \neg B$
2. $F(A, B, C) = A \vee (B \wedge C)$
3. $F(A, B) = A \text{ XOR } B$
4. $F(A, B, C, D) = (A \wedge B) \vee (C \wedge D)$

Exercice N°2 :

Soient les individus suivants :

Individu	A1	A2	Classe
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

Indice d'impureté :

- $Entropie(nœud\ p) = \sum_{k=1}^c P(k|p) * \log_2 (P(k|p))$
- $Gini(nœud\ p) = 1 - \sum_{k=1}^c (P(k|p))^2$

Selon la fonction d'impureté utilisée (entropie, gini) le gain est donné par la fomule :

$$gain(p, A) = i(p) - \sum_{j=1}^n P_j. i(p_j)$$

- Un nœud p
- c le nombre de classe de la variable prédictive
- n le nombre de modalités de la variable A (attribut de test)
- i soit la fonction Entropie ou bien le Gini
- p_j : la proportion des individus du nœud (p) qui vont en position p_j (nœud p_j) selon la valeur de l'attribut A.
- $P(k/p)$ = proportion des individus appartenant à la classe k parmi ceux du nœud p (de la position p)

1. Calculer l'entropie de l'ensemble d'individus par rapport à la valeur de la variable explicative **classe**.
 - a. Calculer l'entropie des deux variables prédictives A1 et A2
 - b. Calculer le gain d'information pour les variables prédictives A1 et A2.
2. Calculer le Gini de l'ensemble d'individus par rapport à la valeur de la variable explicative **classe**.
 - a. Calculer le Gini des deux variables prédictives A1 et A2
 - b. Calculer le gain d'information pour les variables prédictives A1 et A2.

Exercice N°3 :

Soit un échantillon de 200 patients se répartissant en 2 classes : M pour malade et B pour bonne santé. Deux attributs, gorge-irritée et température, permettent de répartir les patients dans chacune des classes suivant le tableau suivant :

	Gorge irritée	Gorge non irritée
Température < 37,5	6B, 37M	91B, 1M
Température ≥ 37,5	2B, 21M	1B, 41M

- Quel(s) arbre(s) de décision peut-on construire à partir de ces données en utilisant le critère du gain d'information (justifier le choix des attributs sans effectuer les calculs).
- On ajoute le critère d'arrêt suivant à l'algorithme de construction de l'arbre : si 90% des individus d'un nœud sont d'une même classe, alors ce nœud devient une feuille de cette classe.
 - Quel arbre obtient-on ?
 - Quel est son taux d'erreur d'apprentissage ?
 - Calculez le taux d'erreur de teste pour l'échantillon de teste suivant :

	Température	Gorge	classe
Patient 1	38	irritée	B
Patient 2	37,2	non irritée	B
Patient 3	37	irrité	M
Patient 4	39	non irritée	M
Patient 5	36,4	irritée	M

Exercice N°2 :

Une compagnie d'assurances automobiles souhaiterait établir des profils parmi ses clients afin de différencier le montant de ses primes. Une première analyse a montré que les attributs qui interviennent le plus sont : le sexe (Homme ou Femme), la couleur du véhicule (Rouge ou Autre), l'âge de l'assuré (Jeune ou Autre), le fait d'avoir des enfants ou non (Oui ou Non).

La variable cible est : Assuré à risque (+) ou non (-).

Les clients ayant souscrits une assurance depuis 5 ans sont analysés selon ces critères. On obtient alors le tableau suivant :

Identificateur	Sexe	Couleur	Age	Enfants	Risque
1	H	A	J	N	+
2	H	R	A	O	-
3	F	R	J	N	+
4	H	R	A	N	+
5	H	R	J	O	+
6	H	A	A	O	-
7	H	A	A	N	-
8	F	A	J	N	-
9	F	A	A	N	-
10	F	R	J	O	-
11	H	R	J	O	+
12	F	A	J	N	-

- Construire l'arbre de décision en utilisant les critères de l'entropie et le gain d'information.
- Construire l'arbre de décision en utilisant le critère de Gini d'information.
- Trouvez quelle classe est attribuée à (sexe=F, Couleur=A, Age=J, Enfant=O) par le classificateur d'arbre de décision.