



Module 7 - Arbres de décisions

Exercices - Consignes

Exercice 1

Quelles sont les différences entre les algorithmes CART, C4.5 et ID3 ?

Exercice 2*

L'objectif de cet exercice est de faire la classification de chiens selon leur race : race pure (Pure) ou race non pure (NPure) (métis). La classification se base sur deux variables Taille et Couleur. La variable Taille peut prendre deux valeurs grand ou petit (G, S respectivement) et la variable Couleur peut également prendre deux valeurs (U) pour uni et (M) pour multiples taches de couleur différente.

Individu	Taille	Couleur	Class
1	P	U	Pure
2	P	U	Pure
3	G	U	Pure
4	G	M	Pure
5	G	M	NPure
6	G	M	NPure
6	P	U	NPure
6	P	M	NPure

1. Calculez l'entropie de la variable Couleur.

2. Calculez l'entropie de la variable Taille.
3. La variable Taille constitue t-elle un bon prédicteur de la race ? Justifiez votre réponse.

Exercice 3*

L'objectif de cet exercice est de refaire les calculs de l'exercice 2 en utilisant le logiciel R.

1. Calculez l'entropie de la variable Couleur.
2. Calculez l'entropie de la variable Taille.
3. Calculez le gain de la variable Couleur et le gain de la variable Taille.

Exercice 4*

L'objectif de cet exercice est de représenter les arbres de décision à partir de deux ensembles de données : les données Iris et les données kyphosis.

1. Installez et chargez les bibliothèques `rpart` et `FSelector`.
2. Importez les deux ensembles de données : `iris` et `kyphosis`.
3. Identifiez les attributs et les attributs cibles de chaque ensemble de données.
4. En utilisant la fonction `rpart` de R, dessinez les arbres de décision correspondant à chaque base de données.
5. Calculez le gain et le rapport de gain pour chacun des attributs.
6. Cherchez l'attribut qui explique le mieux les variables cible `iris$Species` et `kyphosis$Kyphosis`.

7. Redessiner les deux arbres après les éliminations des attributs non importantes en utilisant la fonction `FSelector`. Commentez les résultats.

Exercice 5*

Le service de crédit d'une institution financière désire faire une étude pour classer les clients à fin de déterminer si on peut leur accorder un crédit ou non.

1. Importez l'ensemble de données décrit dans le fichier texte Credit-Data.txt en utilisant R.
2. Divisez l'ensemble de données en deux partitions : les 750 dernières observations pour l'apprentissage et les 250 premières observations le test.
3. En utilisant les deux algorithmes CART et C5.0, représentez les arbres de décision de la base d'apprentissage.
4. Testez les arbres de décision obtenus en utilisant les données de test. Comparez les résultats.