R4.04 : Méthodes d'optimisation
TD 3

Exercice 1 : Arbres de décision

Voici un exemple de jeu de données sur le profilage de personnes ayant acheté un PC. Ce jeu de données contient des informations sur différents facteurs qui pourraient influencer la capacité d'une personne à acheter un ordinateur personnel.

ID	Age	Revenus	Etudiant	Notation crédit	Acheter un PC
1	Jeune	Elevé	Non	Acceptable	Non
2	Jeune	Elevé	Non	Excellent	Non
3	Adulte	Elevé	Non	Acceptable	Oui
4	Senior	Moyen	Non	Acceptable	Oui
5	Senior	Modeste	Oui	Acceptable	Oui
6	Senior	Modeste	Oui	Excellent	Non
7	Adulte	Modeste	Oui	Excellent	Oui
8	Jeune	Moyen	Non	Acceptable	Non
9	Jeune	Modeste	Oui	Acceptable	Oui
10	Senior	Moyen	Oui	Acceptable	Oui
11	Jeune	Moyen	Oui	Excellent	Oui
12	Adulte	Moyen	Non	Excellent	Oui
13	Adulte	Elevé	Oui	Acceptable	Oui
14	Senior	Moyen	Non	Excellent	Non

1. **Quelle est l'entropie initiale du jeu de données ?**
   - Calculer l'entropie initiale pour la variable cible "Acheter un PC" (OUI/NON).
2. **Calculer le gain d'information pour chaque caractéristique possible :**
   - Age
   - Revenue
   - Etudiant (OUI/NON)
   - Notation crédit
3. **Quelle caractéristique offre le plus grand gain d'information ?**
   - Sélectionner la caractéristique qui maximise le gain d'information.
4. **Diviser les données en fonction de cette caractéristique :**
   - Créer deux sous-groupes basés sur la valeur de la caractéristique choisie.
5. **Répéter le processus pour chaque sous-groupe :**
   - Calculer le gain d'information pour chaque caractéristique possible dans chaque sous-groupe.

- Choisissez la caractéristique offrant le plus grand gain d'information pour diviser chaque sous-groupe.

6. **Arrêter le processus :**
   - Arrêter lorsque tous les critères d'arrêt sont atteints (par exemple, toutes les instances dans un nœud sont de la même classe, profondeur maximale de l'arbre atteinte, nombre minimal d'instances dans un nœud).
7. **Construire l'arbre de décision :**
   - Construire l'arbre de manière récursive en suivant les étapes précédentes jusqu'à ce que l'arbre soit complètement construit selon les critères d'arrêt spécifiés.
8. **Quelle décision pour ces nouvelles données :**

Exercice 2 : Algorithme K-means

Soit l'ensemble D des entiers suivants :

D= {2, 5, 8, 10, 11, 18, 20, 13, 4, 6, 12}

On veut répartir les données de D en trois (3) clusters, en utilisant l'algorithme K-means. 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 K-means en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul.

Principe de fonctionnement de l'algorithme K-means:

1. Il faut calculer la distance de chaque valeur de l'ensemble D par rapport à chaque centre de cluster
2. Ajouter la valeur en question au cluster le plus proche.
3. Recalculer à chaque itération les centres de cluster en tant que la moyenne des valeurs y appartenant
4. Re-exécuter les étapes (1) et (2) jusqu'à ce que les valeurs des centres des clusters ne changent plus.
5. Fin de l'algorithme