TD 3 - Méthodes d'Optimisation

Arbre de décision (ID3) et K-means

Exercice 1 : Arbre de décision (ID3)

On dispose du jeu de données suivant :

ID	Age	Revenus	Étudiant	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

La variable cible est **Acheter un PC** (OUI ou NON).

1. Entropie initiale

Comptons le nombre de *OUI* et de *NON* dans la colonne *Acheter un PC* :

OUI : $\{3, 4, 5, 7, 9, 10, 11, 12, 13\}$ (9 occurrences), NON : $\{1, 2, 6, 8, 14\}$ (5 occurrences).

Le total est de 14 instances.

L'entropie est alors :

$$\text{Entropie}(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right).$$

Numériquement,

$$\text{Entropie}(S) \approx 0{,}94.$$

2. Calcul du gain d'information pour chaque caractéristique

(a) Gain d'information : Age

Age prend trois valeurs : **Jeune**, **Adulte**, **Senior**.

— Age = Jeune : IDs $\{1, 2, 8, 9, 11\}$.

$$\text{Acheter=NON} : \{1, 2, 8\} = 3, \quad \text{Acheter=OUI} : \{9, 11\} = 2.$$

Entropie $\approx -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) \approx 0{,}971$.

— Age = Adulte : IDs $\{3, 7, 12, 13\}$.

$$\text{Acheter=OUI} : 4, \quad \text{Acheter=NON} : 0.$$

Entropie = 0 (toutes les instances sont OUI).

— Age = Senior : IDs $\{4, 5, 6, 10, 14\}$.

$$\text{Acheter=OUI} : \{4, 5, 10\} = 3, \quad \text{Acheter=NON} : \{6, 14\} = 2.$$

Même répartition 3/2, Entropie $\approx 0{,}971$.

L'entropie conditionnelle (pondérée) de *Age* :

$$\text{Entropie}_{(\text{Age})} = \frac{5}{14} \times 0{,}971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0{,}971 \approx 0{,}695.$$

Donc,

$$\text{Gain(Age)} = \text{Entropie}(S) - \text{Entropie}_{(\text{Age})} \approx 0{,}94 - 0{,}695 = 0{,}245.$$

(b) Gain d'information : Revenus

Revenus prend trois valeurs : **Elevé**, **Moyen**, **Modeste**.

— Elevé : IDs $\{1, 2, 3, 13\}$. 2 OUI, 2 NON $\implies$ Entropie = 1.

— Moyen : IDs $\{4, 8, 10, 11, 12, 14\}$. 4 OUI, 2 NON $\implies$ Entropie $\approx 0{,}917$.

— Modeste : IDs $\{5, 6, 7, 9\}$. 3 OUI, 1 NON $\implies$ Entropie $\approx 0{,}811$.

Pondération ($|\text{Elevé}| = 4$, $|\text{Moyen}| = 6$, $|\text{Modeste}| = 4$) :

$$\text{Entropie}_{(\text{Revenus})} = \frac{4}{14} \times 1 + \frac{6}{14} \times 0{,}917 + \frac{4}{14} \times 0{,}811 \approx 0{,}286 + 0{,}393 + 0{,}232 = 0{,}911.$$

$$\text{Gain(Revenus)} = 0{,}94 - 0{,}911 = 0{,}029.$$

(c) Gain d'information : Etudiant (Oui/Non)

$$\text{Etudiant} = \text{Oui} : \{5, 6, 7, 9, 10, 11, 13\} \Rightarrow 6 \text{ OUI}, 1 \text{ NON}, \text{ Entropie} \approx 0{,}591,$$
$$\text{Etudiant} = \text{Non} : \{1, 2, 3, 4, 8, 12, 14\} \Rightarrow 3 \text{ OUI}, 4 \text{ NON}, \text{ Entropie} \approx 0{,}985.$$

Les deux sous-groupes ont chacun 7 instances, donc :

$$\text{Entropie}_{(\text{Etudiant})} = 0{,}5 \times 0{,}591 + 0{,}5 \times 0{,}985 = 0{,}296 + 0{,}492 = 0{,}788.$$

$$\text{Gain(Etudiant)} = 0{,}94 - 0{,}788 = 0{,}152.$$

(d) Gain d'information : Notation crédit (Acceptable / Excellent)

$$\text{Acceptable} : \{1,3,4,5,8,9,10,13\} \Rightarrow 6 \text{ OUI}, 2 \text{ NON}, \text{ Entropie} \approx 0{,}812,$$
$$\text{Excellent} : \{2,6,7,11,12,14\} \Rightarrow 3 \text{ OUI}, 3 \text{ NON}, \text{ Entropie} = 1.$$

Pondération :

$$\text{Entropie}_{(\text{Crédit})} = \frac{8}{14} \times 0{,}812 + \frac{6}{14} \times 1 = 0{,}464 + 0{,}429 = 0{,}893.$$
$$\text{Gain}(\text{Crédit}) = 0{,}94 - 0{,}893 = 0{,}047.$$

3. Caractéristique à plus grand gain d'information

On récapitule :

$\text{Gain}(\text{Age}) \approx 0{,}245$, $\text{Gain}(\text{Revenus}) \approx 0{,}029$, $\text{Gain}(\text{Etudiant}) \approx 0{,}152$, $\text{Gain}(\text{Crédit}) \approx 0{,}047$.

C'est *Age* qui maximise le gain d'information ($\approx 0{,}245$) :

On choisit Age comme racine de l'arbre.

4. Diviser les données selon Age

Age=Jeune : $\{1,2,8,9,11\}$
Age=Adulte : $\{3,7,12,13\}$
Age=Senior : $\{4,5,6,10,14\}$

5. Répéter le processus

Sous-groupe (Age=Jeune) : 5 instances $\{1,2,8,9,11\}$. Répartition (3 NON, 2 OUI).
Vérifions l'attribut *Etudiant* :

$$\text{Etudiant=Oui} : \{9,11\} \Rightarrow (2 \text{ OUI}, 0 \text{ NON}),$$
$$\text{Etudiant=Non} : \{1,2,8\} \Rightarrow (0 \text{ OUI}, 3 \text{ NON}).$$

Chacun des deux sous-groupes est **pur** $\Longrightarrow$ entropie conditionnelle = 0.

$$\text{Gain} = \approx 0{,}971.$$

On choisit *Etudiant* pour séparer Jeune.

Sous-groupe (Age=Adulte) : $\{3,7,12,13\}$, tous OUI (4 OUI, 0 NON). Entropie=0.

Feuille : Décision = OUI pour Age=Adulte.

Sous-groupe (Age=Senior) : $\{4,5,6,10,14\}$, répartition (3 OUI, 2 NON). Entropie $\approx 0{,}971$.
Vérifions *Notation crédit* :

$$\text{Acceptable} : \{4,5,10\} \Rightarrow (3 \text{ OUI}, 0 \text{ NON}),$$
$$\text{Excellent} : \{6,14\} \Rightarrow (0 \text{ OUI}, 2 \text{ NON}).$$

Entropie conditionnelle=0, gain=0,971.

On sépare Senior par Notation crédit.

6. Arrêter

Tous les nœuds fils sont purs ou n'ont plus d'attribut utile. On s'arrête.

7. L'arbre de décision

On obtient la structure ci-dessous (présentation textuelle) :

Racine : Age
Jeune → (étudiant ?)
Étudiant=Non ⇒ Acheter=Non
Étudiant=Oui ⇒ Acheter=Oui

Adulte ⇒ Acheter=Oui (feuille)

Senior → (notation crédit ?)
Acceptable ⇒ Acheter=Oui
Excellent ⇒ Acheter=Non

8. Décision pour de nouvelles données

Pour classifier un nouveau tuple (Age, Revenus, Etudiant, Crédit) :
— Commencer par la racine *Age*.
— Suivre la branche correspondante (Jeune, Adulte ou Senior).
— Poser la question sur *Etudiant* (si *Jeune*) ou sur *Crédit* (si *Senior*).
— Arriver à la feuille qui donne la décision OUI ou NON.

Exercice 2 : Algorithme K-means

Soit l'ensemble

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

On veut former **3** clusters (k=3) en utilisant l'algorithme K-means, avec pour centres initiaux :

$$C_1^{(0)} = 8, \quad C_2^{(0)} = 10, \quad C_3^{(0)} = 11.$$

La distance est $d(a, b) = |\, a - b\,|$.

Principe K-means (rappel)

1. Calculer la distance de chaque point aux centres de cluster.
2. Affecter chaque point au cluster du centre le plus proche.
3. Recalculer le centre de chaque cluster comme la *moyenne* des points qui y sont affectés.
4. Répéter tant que les centres changent.

Itération 1

$$C_1^{(0)} = 8, \quad C_2^{(0)} = 10, \quad C_3^{(0)} = 11.$$

On calcule la distance de chaque $x \in D$:

Point	$\lvert x-8\rvert$	$\lvert x-10\rvert$	$\lvert x-11\rvert$	**Cluster**
2	6	8	9	C1 (6)
5	3	5	6	C1 (3)
8	0	2	3	C1 (0)
10	2	0	1	C2 (0)
11	3	1	0	C3 (0)
18	10	8	7	C3 (7)
20	12	10	9	C3 (9)
13	5	3	2	C3 (2)
4	4	6	7	C1 (4)
6	2	4	5	C1 (2)
12	4	2	1	C3 (1)

$$C_1^{(1)} = \{2, 5, 8, 4, 6\},$$
$$C_2^{(1)} = \{10\},$$
$$C_3^{(1)} = \{11, 18, 20, 13, 12\}.$$

Recalcule des centres (moyennes) :

$$C_1 \leftarrow \frac{2+5+8+4+6}{5} = 5,$$
$$C_2 \leftarrow 10,$$
$$C_3 \leftarrow \frac{11+18+20+13+12}{5} = \frac{74}{5} = 14{,}8 \approx 15.$$

Nouveaux centres :

$$C_1^{(1)} = 5, \quad C_2^{(1)} = 10, \quad C_3^{(1)} \approx 15.$$

Itération 2

Centres : 5, 10, 15. On refait l'affectation :

Point	$\lvert x-5\rvert$	$\lvert x-10\rvert$	$\lvert x-15\rvert$	**Cluster**
2	3	8	13	C1
5	0	5	10	C1
8	3	2	7	C2
10	5	0	5	C2
11	6	1	4	C2
18	13	8	3	C3
20	15	10	5	C3
13	8	3	2	C3
4	1	6	11	C1
6	1	4	9	C1
12	7	2	3	C2

Nouveaux clusters :

$$\begin{aligned} C_1 &= \{2, 5, 4, 6\}, \\ C_2 &= \{8, 10, 11, 12\}, \\ C_3 &= \{18, 20, 13\}. \end{aligned}$$

Recalcul des centres :

$$\begin{aligned} C_1 &\leftarrow \frac{2+5+4+6}{4} = 4{,}25, \\ C_2 &\leftarrow \frac{8+10+11+12}{4} = 10{,}25, \\ C_3 &\leftarrow \frac{18+20+13}{3} = 17. \end{aligned}$$

Itération 3

Centres : 4,25, 10,25, 17. On réaffecte et on aboutit à des sous-ensembles stables (détails omis pour la distance). Finalement, on obtient :

$$C_1 = \{2, 4, 5, 6\}, \quad C_2 = \{8, 10, 11, 12, 13\}, \quad C_3 = \{18, 20\}.$$

Nouveaux centres :

$$\begin{aligned} C_1 &= 4{,}25, \\ C_2 &= \frac{8+10+11+12+13}{5} = 10{,}8, \\ C_3 &= \frac{18+20}{2} = 19. \end{aligned}$$

$\Longrightarrow$ A l'itération suivante, plus aucun changement. L'algorithme s'arrête.

Résultat final

Les trois clusters finaux sont :

$$\boxed{\begin{array}{l}\text{Cluster } 1 = \{2, 4, 5, 6\}, \quad \text{centre} \approx 4{,}25, \\ \text{Cluster } 2 = \{8, 10, 11, 12, 13\}, \quad \text{centre} \approx 10{,}8, \\ \text{Cluster } 3 = \{18, 20\}, \quad \text{centre} = 19.\end{array}}$$