Méthode des arbres de décision

Théorie, démonstrations et interprétation des résultats

1 Introduction

Les arbres de décision sont des méthodes d'apprentissage supervisé utilisées en classification et en régression. Ils reposent sur une partition récursive de l'espace des variables explicatives afin de produire un modèle interprétable sous forme de règles logiques.

2 Formulation mathématique du modèle

Soit un échantillon d'apprentissage :

$$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n},$$

où $X_i \in \mathbb{R}^p$ est un vecteur de variables explicatives et Y_i la variable cible.

Un arbre de décision définit une partition de l'espace $\mathbb{R}^p$ en régions disjointes $\{R_1, \ldots, R_M\}$, et une fonction prédictive :

$$f(X) = \sum_{m=1}^{M} c_m \mathbb{1}_{X \in R_m},$$

où c_m est une constante associée à la région R_m.

3 Construction récursive de l'arbre

À chaque nœud, on cherche une coupure de la forme :

$$X_j \leq s \quad \text{ou} \quad X_j > s,$$

où $j \in \{1, \ldots, p\}$ et $s \in \mathbb{R}$.

Le choix optimal repose sur la minimisation d'une fonction d'impureté.

4 Critères d'impureté

4.1 Classification

Soit K le nombre de classes et p_k la proportion d'observations appartenant à la classe k dans un nœud t.

4.1.1 Indice de Gini

L'impureté de Gini d'un nœud t est définie par :

$$G(t) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2.$$

Un nœud est dit *pur* si toutes les observations appartiennent à une seule classe, ce qui équivaut à $G(t) = 0$.

4.1.2 Moyenne pondérée de l'impureté de Gini

Lorsqu'un nœud parent t est séparé en deux sous-nœuds t_g et t_d, ces sous-nœuds peuvent contenir des effectifs différents. Il est donc nécessaire de prendre en compte leur taille relative afin d'évaluer correctement la qualité d'une coupure.

Après une coupure, l'impureté globale est mesurée par la **moyenne pondérée de l'impureté de Gini** :

$$G_{\text{pondéré}} = \frac{n_g}{n} G(t_g) + \frac{n_d}{n} G(t_d),$$

où :

- n est le nombre total d'observations dans le nœud parent,
- n_g et n_d sont les effectifs des sous-nœuds gauche et droit.

Cette pondération permet de donner plus d'importance aux sous-nœuds contenant un grand nombre d'observations et évite qu'un sous-nœud de petite taille mais très pur influence excessivement le choix de la coupure.

La coupure optimale est celle qui **minimise** la moyenne pondérée de l'impureté de Gini, ou de manière équivalente, celle qui **maximise** la diminution d'impureté :

$$\Delta G = G(t) - G_{\text{pondéré}}.$$

4.1.3 Entropie de Shannon

Une autre mesure de l'impureté d'un nœud t est l'entropie de Shannon :

$$H(t) = -\sum_{k=1}^{K} p_k \log(p_k).$$

Le gain d'information associé à une coupure est défini par :

$$\Delta H = H(t) - \left(\frac{n_g}{n} H(t_g) + \frac{n_d}{n} H(t_d) \right).$$

La coupure retenue est celle qui maximise le gain d'information.

4.2 Régression

On minimise la somme des carrés des résidus :

$$\mathrm{RSS}(t) = \sum_{i \in t} (Y_i - c)^2$$

Proposition 1. *La valeur optimale de c minimisant $RSS(t)$ est la moyenne empirique :*

$$c_t = \frac{1}{|t|} \sum_{i \in t} Y_i$$

Proof. On dérive par rapport à c :

$$\frac{d}{dc} \sum (Y_i - c)^2 = -2 \sum (Y_i - c)$$

L'annulation de la dérivée donne :

$$c = \frac{1}{|t|} \sum Y_i$$

□

5 Justification théorique

5.1 Arbres comme estimateurs non paramétriques

Les arbres de décision sont des estimateurs par morceaux constants. Lorsque la profondeur augmente, ils peuvent approximer toute fonction mesurable, au prix d'une variance plus élevée.

5.2 Compromis biais–variance

- Arbre profond : biais faible, variance élevée
- Arbre peu profond : biais élevé, variance faible

6 Élagage (Pruning)

On introduit un critère pénalisé :

$$R_\alpha(T) = R(T) + \alpha |T|$$

où $|T|$ est le nombre de feuilles de l'arbre et $\alpha > 0$ un paramètre de complexité.

Théorème 1. *Pour toute valeur de α, il existe un sous-arbre optimal minimisant $R_\alpha(T)$.*

7 Interprétation des résultats

7.1 Règles de décision

Chaque chemin racine-feuille correspond à une règle logique :

$$(X_1 \leq s_1) \wedge (X_2 > s_2) \Rightarrow \text{Classe } k$$

7.2 Importance des variables

L'importance d'une variable X_j est définie par :

$$\text{Imp}(X_j) = \sum_{t \in \mathcal{T}_j} \Delta I(t)$$

où $\Delta I(t)$ est la diminution d'impureté au nœud t.

8 Limites théoriques

- Instabilité vis-à-vis des données
- Coupures axis-parallèles uniquement
- Optimisation locale (méthode gloutonne)

9 Conclusion

Les arbres de décision offrent un compromis intéressant entre performance prédictive et interprétabilité, mais nécessitent un contrôle de la complexité pour éviter le sur-apprentissage.