

Arbres de décision

Séance 4 – CART et coût empirique régularisé

Christophe Ambroise 

UEVE, UMR CNRS 8071

2 octobre 2026

1

Changement de terrain

2

Un nouveau jeu de données

Adult – recensement américain de 1994.

Cible : revenu annuel supérieur à 50 000 dollars.

10 000 observations, 12 variables explicatives, 24 % de positifs.

4 variables numériques : âge, heures travaillées...

8 variables catégorielles : emploi, diplôme, statut marital, profession, situation familiale, origine, sexe, pays de naissance.

3

Ce qui change

Les modèles gaussiens supposent des variables continues. Beaucoup de variables d'**Adult** sont catégorielles.

Encoder les professions par des nombres impose un ordre arbitraire.

Une modalité comme étiquette, pas comme quantité.

4

Des valeurs manquantes

Valeurs manquantes dans `workclass`, `occupation` et `native-country`.

Manquants non aléatoires. `occupation` manque surtout chez les personnes sans emploi.

Une valeur manquante peut être informative. Une modalité dédiée conserve cette information.

5

Partitionnement récursif

6

L'idée

Des régions rectangulaires. Une prédiction constante par région.

Une seule variable par découpe : $x_j \leq s$, ou $x_j \in A$ pour une variable catégorielle.

Frontières parallèles aux axes. Une contrainte forte, source de lisibilité.

7

Partition et arbre

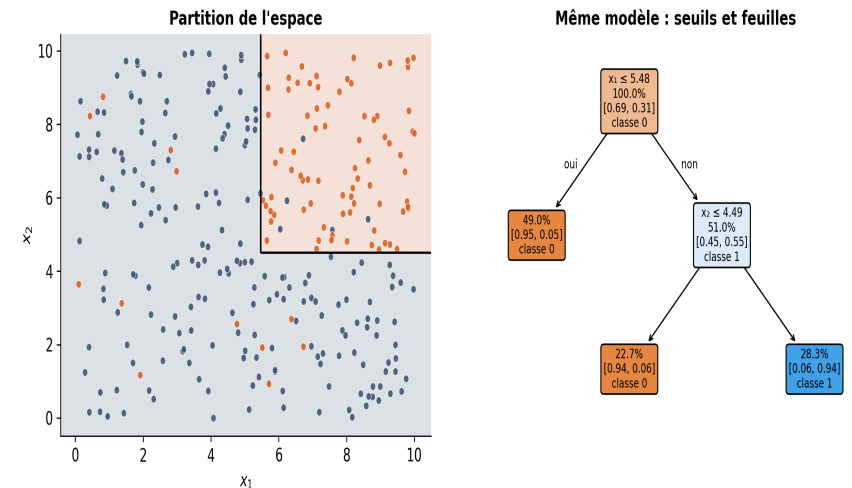


Figure 1

Un modèle, deux lectures. Chaque nœud interne correspond à une frontière noire.

8

Choisir une découpe

Au nœud t , CART choisit le couple (j, s) qui maximise la baisse d'impureté :

$$\Delta i(j, s) = i(t) - \frac{n_g}{n_t} i(t_g) - \frac{n_d}{n_t} i(t_d)$$

Recherche exhaustive sur les variables et les seuils.

Choix **glouton** : optimum local, sans garantie globale. Aucun retour en arrière.

9

Mesurer l'impureté

Pour une proportion p de la classe 1 dans un nœud :

Gini	$i(t) = 2p(1 - p)$
entropie	$i(t) = -p \log_2 p - (1 - p) \log_2 (1 - p)$
erreur	$i(t) = \min(p, 1 - p)$

10

Une découpe, chiffrée

Nœud de 100 personnes : 40 % au-dessus de 50 k.i(t) = 2 4 6 = 0{,}48\$.

Découpe sur **marital-status** :

	n	p	Gini
marié	50	0,64	0,461
non marié	50	0,16	0,269

$$\Delta i = 0,48 - \frac{50}{100} (0,461) - \frac{50}{100} (0,269) = 0,115$$

Comparaison avec toutes les autres découpes.

11

Pourquoi pas le taux d'erreur

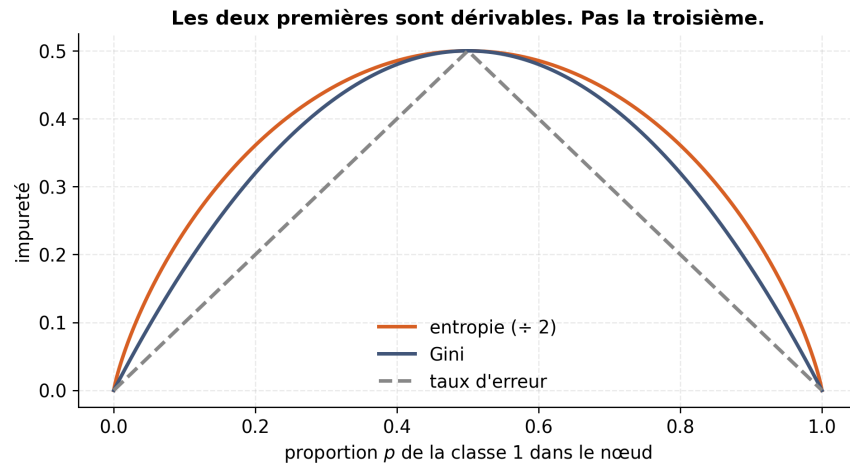


Figure 2

12

La raison est géométrique

Le taux d'erreur est linéaire par morceaux. Son gain peut rester **nul** malgré deux fils plus purs.

Gini et entropie sont strictement concaves. Toute séparation utile réduit l'impureté.

Gini et entropie pour **faire croître** l'arbre. Le taux d'erreur pour **l'élager**.

13

Quand s'arrêter

Pas de critère d'arrêt local fiable.

Un seuil sur Δi peut arrêter trop tôt. Un faible gain peut ouvrir sur de bonnes découpes.

Réponse de Breiman : arbre maximal, puis élagage.

14

L'arbre maximal sur-apprend

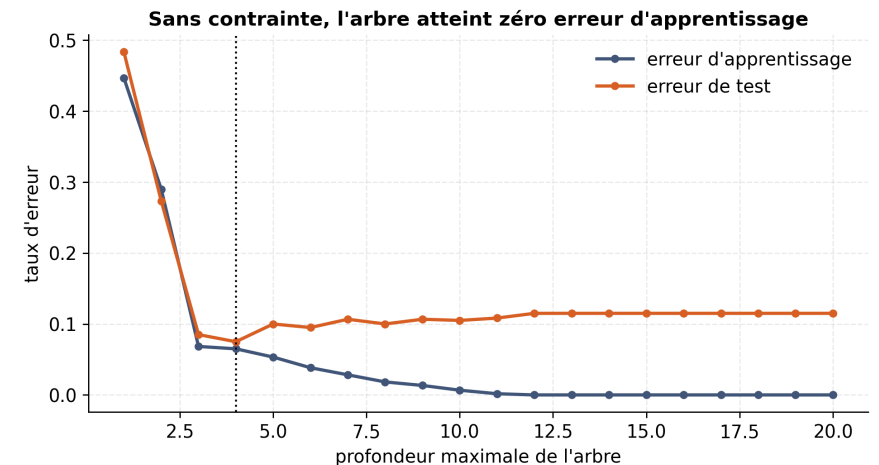


Figure 3

Une feuille par observation : erreur d'apprentissage nulle, erreur de test en hausse.

15

Le coût pénalisé

16

Le principe, rappelé

Séance 1 :

$$f^* = \arg \min_{f \in \mathcal{F}} \hat{R}(f, D_n) + H(f)$$

avec H croissant avec la capacité de f .

CART : une réalisation exacte de ce principe.

17

Le coût-complexité de Breiman

Pour un arbre T , de $|T|$ feuilles :

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

$R(T)$: taux d'erreur d'apprentissage, décroissant avec la taille.

$|T|$: nombre de feuilles, donc **capacité**.

$\alpha \geq 0$: prix d'une feuille.

18

Deux cas limites

$\alpha = 0$ – aucune pénalité, arbre maximal.

α très grand – arbre réduit à sa racine.

Entre les deux : arbre de taille intermédiaire.

19

Le résultat de Breiman

Pour chaque α , il existe un **unique** plus petit sous-arbre T_α qui minimise R_α .

Quand α croît de 0 à l'infini, ces T_α forment une suite **finie** et **emboîtée** :

$$T_{\max} \supset T_1 \supset T_2 \supset \dots \supset \{\text{racine}\}$$

Un continuum réduit à une liste finie.

20

Le maillon faible

Pour un nœud interne t , soit T_t la branche qui en part. On compare garder la branche ou la remplacer par une feuille :

$$g(t) = \frac{R(t) - R(T_t)}{|T_t| - 1}$$

$g(t)$: baisse d'erreur par feuille supplémentaire.

Élagage du g minimal – le **maillon faible**. Répétition jusqu'à la racine.

21

Le chemin d'élagage

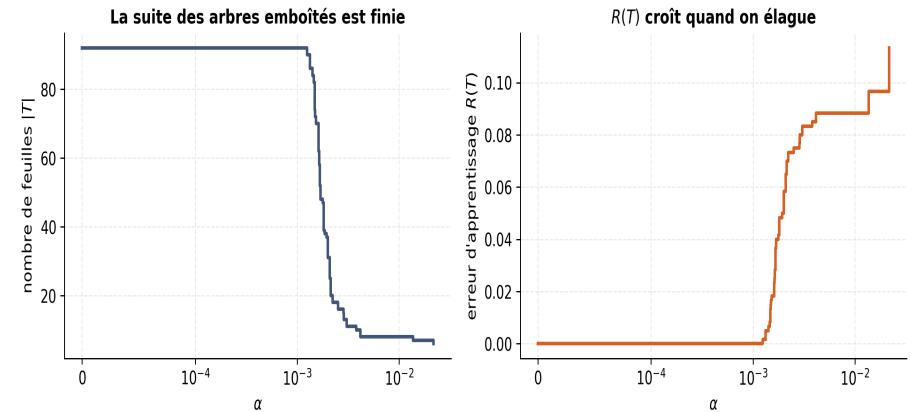


Figure 4

22

Choisir α

α : hyperparamètre choisi par validation croisée.

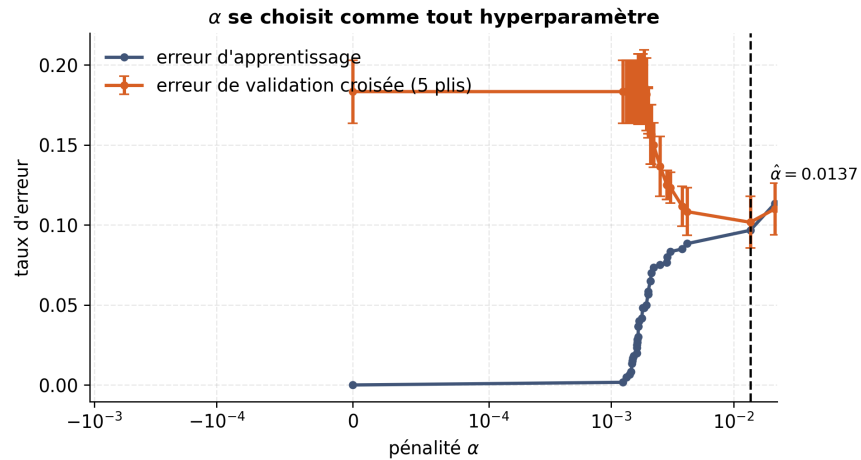


Figure 5

Ce que la figure montre

L'erreur d'apprentissage croît avec α . Inutile pour choisir.

Minimum **intérieur** de l'erreur de validation : $\hat{\alpha}$.

Compromis biais-variance réglé par un seul nombre.

En scikit-learn

```
1 from sklearn.tree import DecisionTreeClassifier
2
3 # la suite emboîtée et ses alpha de coupure
4 complet = DecisionTreeClassifier(random_state=0).fit(X, y)
5 chemin = complet.cost_complexity_pruning_path(X, y)
6 chemin.ccp_alphas
7
8 # un arbre élagué pour un alpha donné
9 arbre = DecisionTreeClassifier(ccp_alpha=0.002).fit(X, y)
```

`cost_complexity_pruning_path` implémente le maillon faible.

Un arbre, en entier

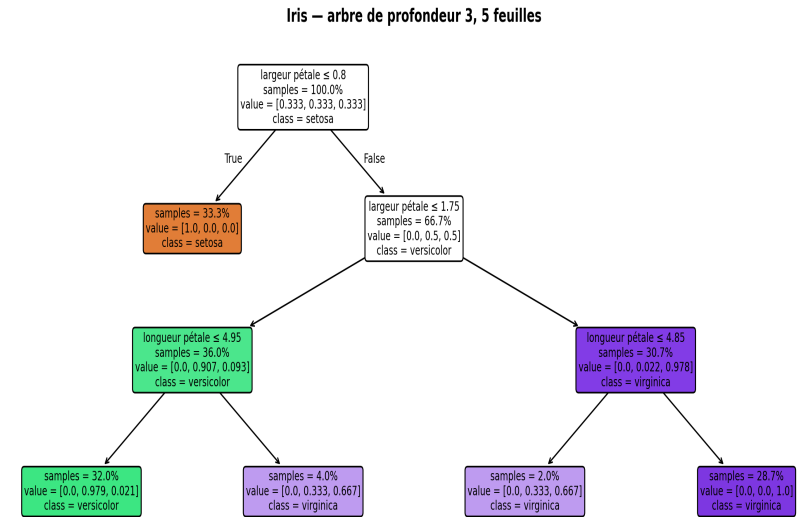


Figure 6

Ce que cet arbre dit

Première découpe : situation familiale. Puis diplôme et revenus du capital.

LDA fournit huit coefficients à lire ensemble. L'arbre fournit des règles.

Interprétabilité par la **forme** du modèle. Pas d'explication ajoutée après coup.

Compléments

Arbres de régression

Même construction. Deux substitutions.

Impureté : variance intra-nœud. $i(t) = \frac{1}{n_t} \sum_{i \in t} (y_i - \bar{y}_t)^2$.

Prédiction d'une feuille : moyenne des y_i .

Fonction constante par morceaux. Aucune extrapolation.

29

Importance des variables

Pour chaque variable j : somme pondérée des baisses d'impureté.

Biais envers les variables à **nombreuses modalités** :
davantage de découpes, donc davantage de gains fortuits.

Dans **Adult**, **sex** et **race** demandent une prudence supplémentaire.

30

Ce que scikit-learn n'implémente pas

CART original : **bipartition des modalités** en deux groupes.

Valeurs manquantes : **scissions de substitution** fondées sur d'autres variables.

scikit-learn n'implémente aucune des deux. Encodage numérique obligatoire.

31

Autres familles d'arbres

- **ID3** et **C4.5** — Quinlan. Entropie, gain ratio, scissions multiples.
- **CHAID** — Kass. Fusion et découpe guidées par un test du χ^2 .
- **Arbres d'inférence conditionnelle** — Hothorn et al. Tests statistiques, sans biais de sélection.

Même partitionnement récursif. Spécificité de CART : arbre maximal, puis élagage pénalisé.

32

Forces

- Interprétable : prédiction sous forme de conditions.
- Invariant aux transformations monotones de chaque variable.
- Variables de natures différentes dans la formulation originale.
- Sélectionne les variables au passage.

33

Limites

- Frontières parallèles aux axes. Une séparation oblique demande un escalier.
- **Instable** : quelques observations peuvent modifier tout l'arbre.
- Prédiction constante par morceaux, sans extrapolation.

34

L'instabilité, montrée

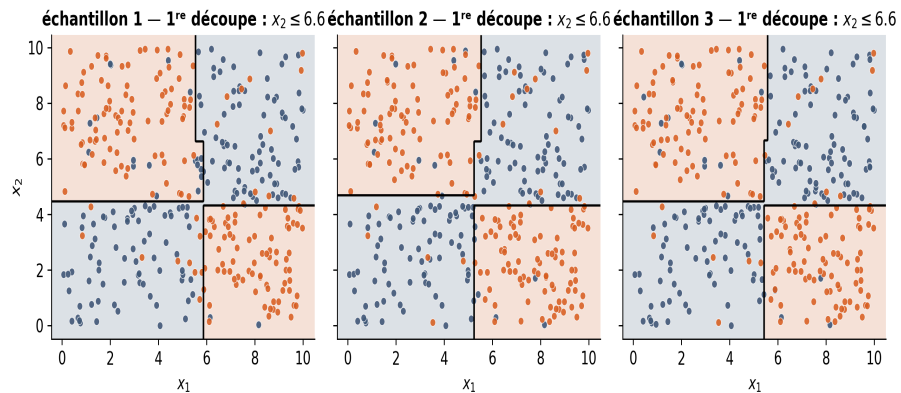


Figure 7

Trois sous-échantillons à 90 %. Trois arbres différents.

Instabilité : variance élevée.

35

En bref

1. Découpes parallèles aux axes, choisies gloutonnement.
2. Gini et entropie font croître l'arbre ; le taux d'erreur ne le peut pas.
3. Faire croître au maximum, puis élaguer.
4. $R_\alpha(T) = R(T) + \alpha|T|$ est le risque empirique pénalisé, sous forme d'une suite finie d'arbres emboîtés.
5. $\hat{\alpha}$ se choisit par validation croisée.