Arbres CART

UE IME204, M2 SITIS, ISPED
UE STA307, M2 EPI, ISPED

Justine Remiat, support de cours Robin Genuer

26/01/2024

Arbres CART

Introduction

- Les arbres CART (Classification And Regression Trees) ont été introduits par Breiman, Friedman, Olshen & Stone (1984)
- Ils font partie de la famille des méthodes d'arbres de décision, introduite depuis les années 70
- Le gros avantage des arbres CART est qu'ils proposent un moyen de régler "automatiquement" la taille de l'arbre (voir la partie Élagage plus loin)
- De plus, c'est un algorithme qui est la base de méthodes très performantes (Boosting, Bagging, RF, ...)

Leo Breiman

- De CART aux Forêts Aléatoires (RF) : 20 ans d'une trajectoire scientifique
- Travaux initialement en probabilités, il a ensuite marqué de son empreinte la statistique appliquée et l'apprentissage statistique
- Série de papiers fondamentaux dans *Annals of Statistics* et *Machine Learning*

Cadre

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ v.a. i.i.d. de même loi que (X, Y).

$X \in \mathbb{R}^p$ (p variables d'entrée quantitatives, mais la méthode gère facilement des variables qualitatives)

$Y \in \mathcal{Y}$ (variable de sortie ou réponse) :

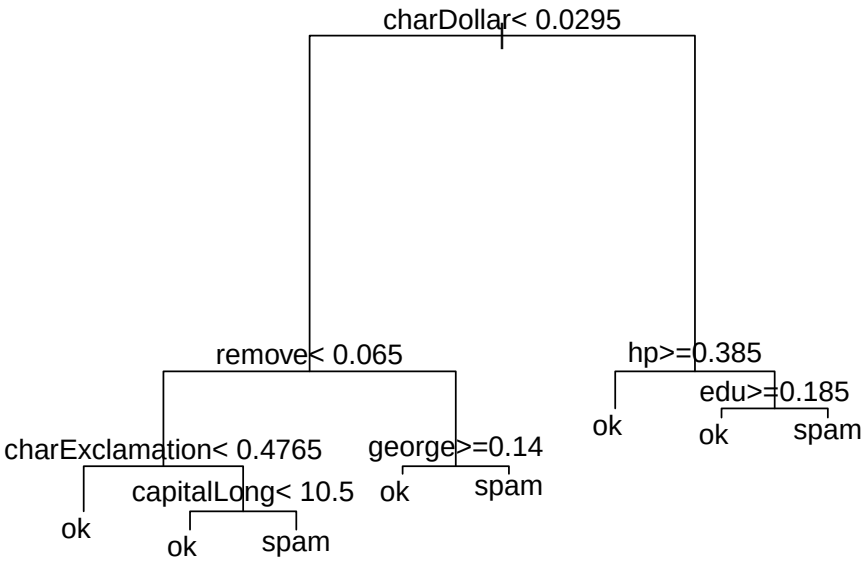
- $\mathcal{Y} = \mathbb{R}$: régression
- $\mathcal{Y} = \{1, \ldots, L\}$: classification

But : construire un prédicteur $\widehat{h} : \mathbb{R}^p \to \mathcal{Y}$

Le jeu de données spam

- But : Construire un détecteur automatique de spams et déterminer les variables importantes
- $n = 4601$ emails (1813 spams, $\approx 40\%$)
- $p = 57$ prédicteurs:
  - 54 sont des proportions de mots ou de caractères comme *charDollar ($)*, *charExclamation (!)*, *remove*, *free*
  - 2 liées aux longueurs de suites de majuscules (moyenne, maximum) et enfin le nombre de majuscules

Un arbre CART pour les données `spam`



Structure de l'arbre : 7 noeuds internes et 8 feuilles; splits basés sur *charDollar*, *remove*, *hp*, *charExclamation*, *george*, *edu* et *capitalLong*

Prédiction par l'arbre : les feuilles donnent les prédictions de Y (*spam* ou *ok*) et sa distribution

Interprétation : de la racine à la feuille la plus à droite : si "beaucoup" de *charDollar*, "peu" de *hp* et "peu" de *edu*, alors "presque toujours" spam

Principe

Arbre : prédicteur constant par morceaux, obtenu par partitionnement récursif binaire de $\mathbb{R}^p$

Restriction : coupures parallèles aux axes

Classiquement, à chaque étape du partitionnement, on vise à séparer "au mieux" les données du noeud courant, en deux noeuds fils

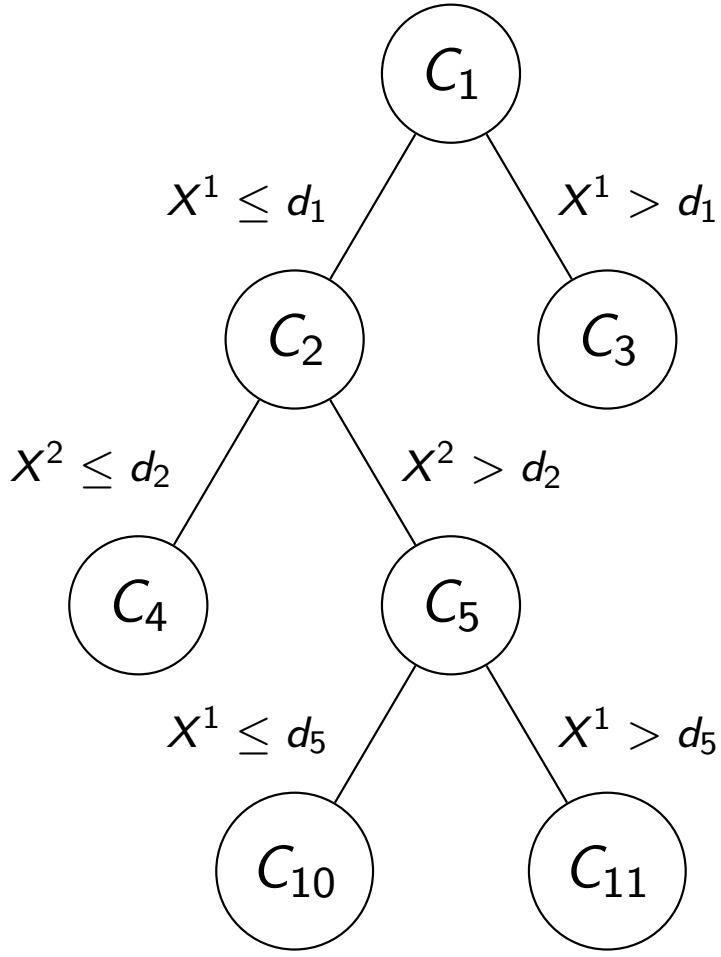


Figure 1: Arbre de classification

Arbre CART et fonction constante par morceaux

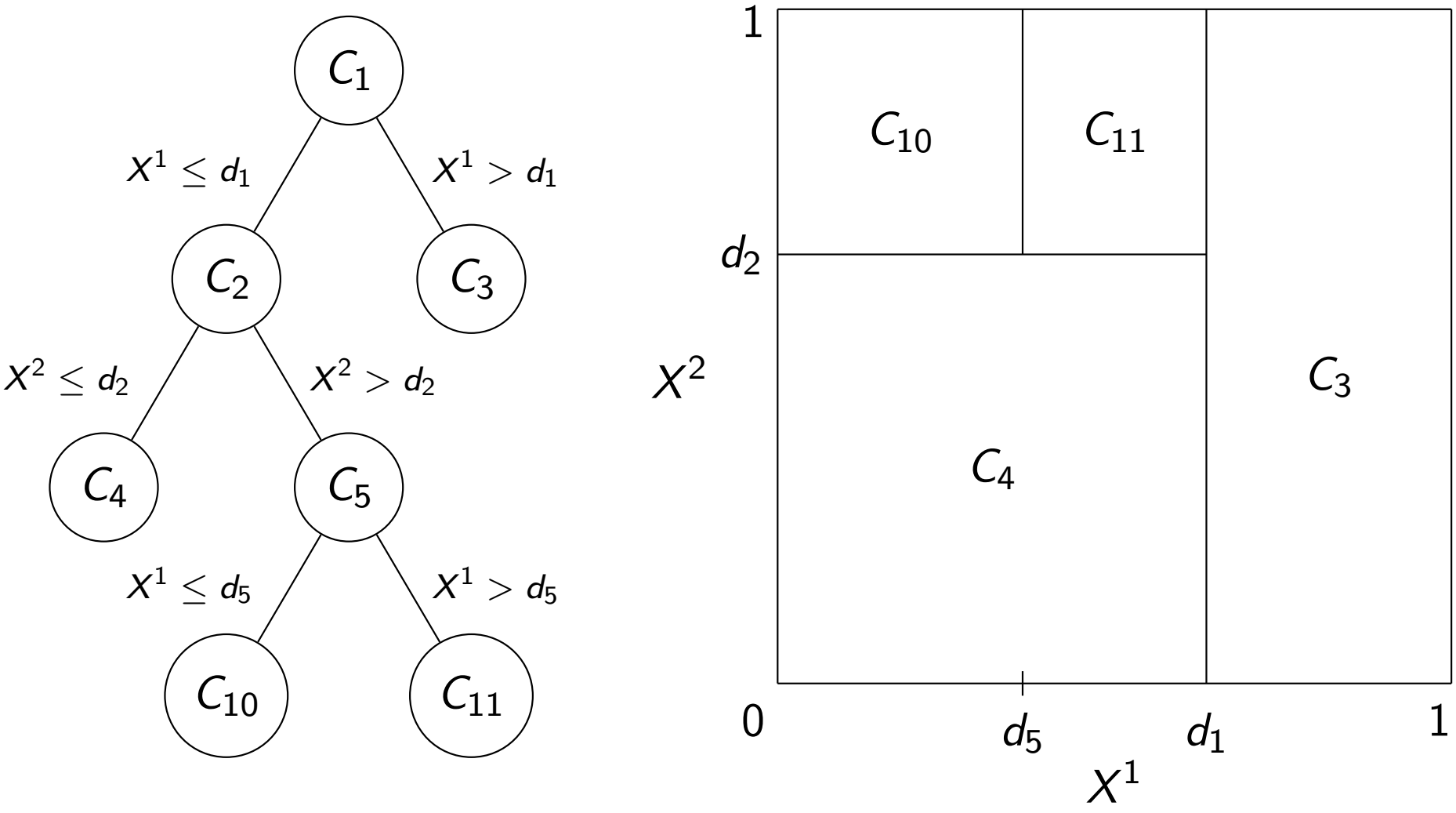


Figure 2: Arbre de classification et partition associée

Arbre de classification *v.s* arbre de régression

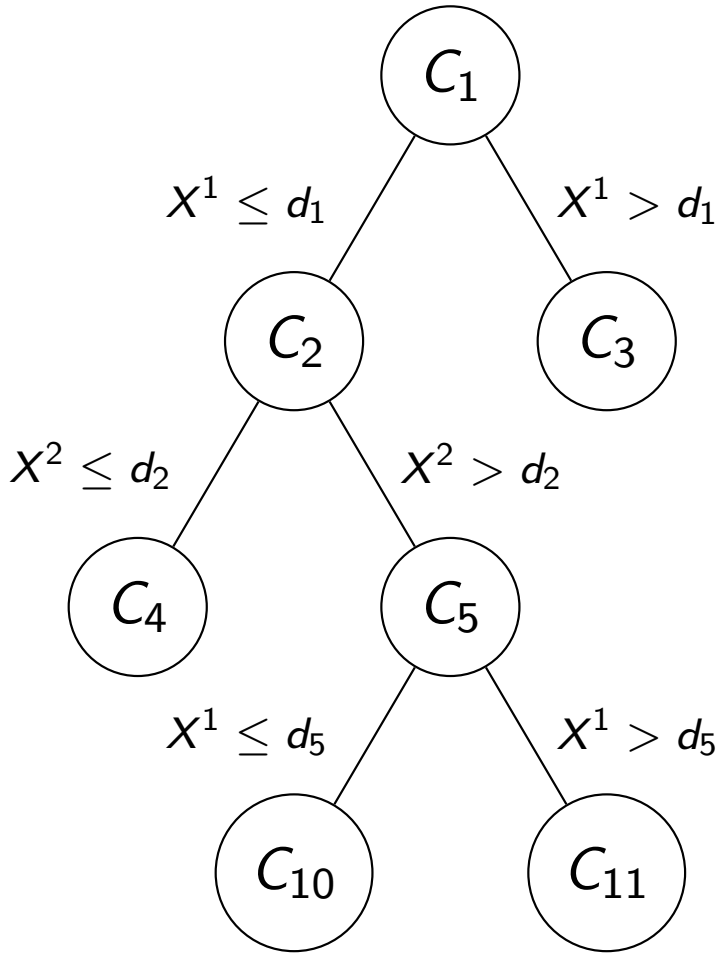


Figure 3: Arbre de classification

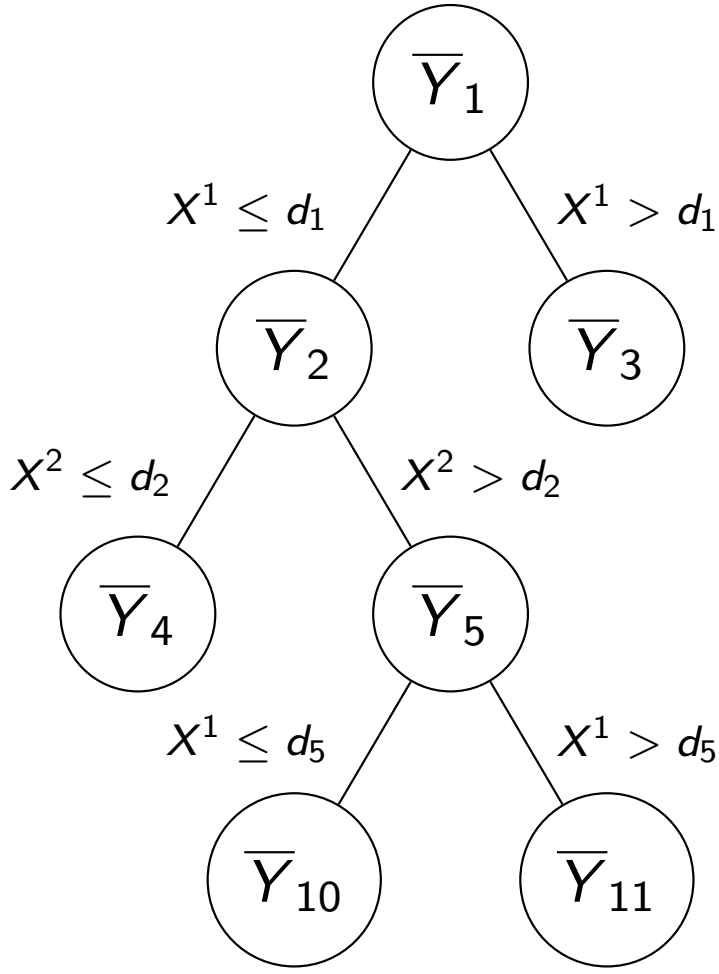


Figure 4: Arbre de régression

Construction

- Découpe (ou coupure ou split) :
  - X^j quanti : $\{X^j \leq d\} \cup \{X^j > d\}$ avec d entre le min et le max de X^j
  - X^j quali : $\{X^j \in A\} \cup \{X^j \in \overline{A}\}$ avec A un sous-ensemble des modalités de X^j
- Hétérogénéité :
  - Régression : $\Phi(t) = \frac{1}{\#t} \sum_{i:x_i \in t} (y_i - \overline{y}_t)^2$ variance d'un nœud t
  - Classification : $\Phi(t) = \sum_{c=1}^{L} \hat{p}_t^c (1 - \hat{p}_t^c)$, indice de Gini d'un nœud t, où $\hat{p}_t^c$ est la proportion d'observations de classe c dans le nœud t.
- Optimisation : on maximise :

$$\Phi(t) - \left(\frac{\#t_L}{\#t} \Phi(t_L) + \frac{\#t_R}{\#t} \Phi(t_R) \right)$$

Arbre maximal et élagage

Arbre maximal, règle d'arrêt:

- Partitionnement récursif par maximisation locale de la décroissance de l'hétérogénéité
- Ne pas découper un noeud pur ou à trop faible effectif

Elagage:

- Arbre maximal T_{max} sur-ajusté aux données : sur-apprentissage
- On cherche un arbre optimal qui est un sous-arbre élagué de T_{max} minimisant le critère pénalisé :

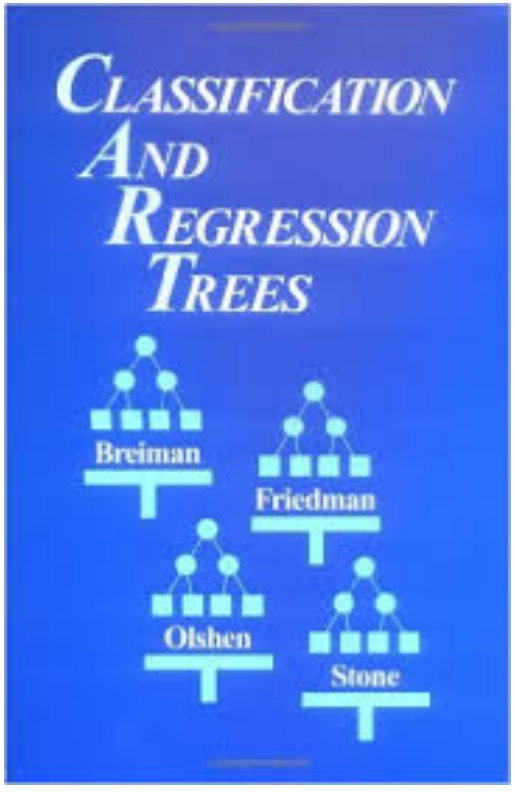
$$crit_\alpha(T) = R_{\text{emp}}(\hat{h}_T) + \alpha \frac{|T|}{n}$$

avec $R_{\text{emp}}(\hat{h}_T)$ l'erreur empirique du prédicteur associé à T et $|T|$ le nombre de feuilles de T

Algorithme d'élagage

Entrée	Arbre maximal T_{max}
Initialisation	$\alpha_1 = 0$, $T_1 = T_{\alpha_1} = \underset{T \text{ élagué de } T_{max}}{\arg\min} \ \overline{err}(T)$ `initialiser` $T = T_1$ et $k = 1$
Iteration	`Tant que` $\lvert T \rvert > 1$, `Calculer` $\alpha_{k+1} = \min_{\{t \text{ noeud interne de } T\}} \frac{\overline{err}(t) - \overline{err}(T_t)}{\lvert T_t \rvert - 1}$ `Elaguer` toutes les branches T_t de T telles que $\overline{err}(T_t) + \alpha_{k+1}\lvert T_t \rvert = \overline{err}(t) + \alpha_{k+1}$ `Prendre` T_{k+1} le sous-arbre élagué ainsi obtenu. `Boucler sur` $T = T_{k+1}$ et $k = k + 1$
Sortie	Arbres $T_1 \succ \ldots \succ T_K = \{t_1\}$ Paramètres $(\alpha_1, \ldots, \alpha_K)$

Références



- CART Classification And Regression Trees, Breiman et al. (1984)
- Présentation en français de CART en régression dans le chapitre 2 de la thèse S. Gey (2002)
- Voir aussi Zhang, Singer (2010) et Hastie, Tibshirani, Friedman (2009)

Extensions ou variantes

- Variantes
  - En régression, prédicteurs plus réguliers, comme MARS Friedman (1991)
  - Ortho-CART Donoho et al. (1997) : en traitement d'images, splits dyadiques + élagage
  - Dyadic-CART : généralisation dans Blanchard et al. (2007)
- Extensions
  - Données de survie LeBlanc & Crowley (1993), Molinaro et al. (2004), Bou-Hamad et al. (2011)
  - Données spatiales Bel et al. (2009)
  - Données longitudinales ou fonctionnelles Zhang, Singer (2010), Captaine et al. (2020)

Discussion

Avantages

- Le principal : interprétabilité
- Méthode non-paramétrique, pas d'hypothèse sur les données
- Cadre unique pour régression et classification (binaire ou multi-classes) + entrées quantitatives et/ou qualitatives
- Découpes compétitives : développement manuel de l'arbre et importance des variables
- Traitement élégant des valeurs manquantes en prédiction : découpes de substitution

Inconvénient

- Le principal : instabilité
- Non-paramétrique $\Longrightarrow$ pas d'estimation, pas d'IC, pas de test