

# Arbres de régression et de classification

Véronique Tremblay

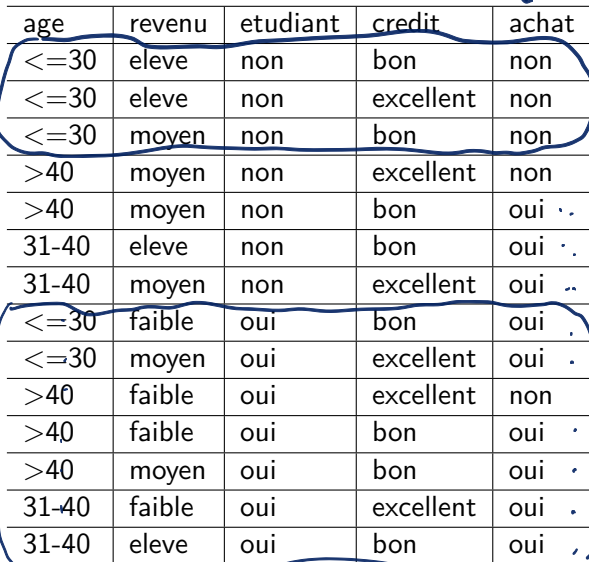
# Introduction

# Objectifs

- Comprendre le concept de base
- Connaître le vocabulaire

$n=14$ 

Exemple



age	revenu	etudiant	credit	achat
$\leq 30$	eleve	non	bon	non
$\leq 30$	eleve	non	excellent	non
$\leq 30$	moyen	non	bon	non
$> 40$	moyen	non	excellent	non
$> 40$	moyen	non	bon	oui .
31-40	eleve	non	bon	oui .
31-40	moyen	non	excellent	oui ..
$\leq 30$	faible	oui	bon	oui .
$\leq 30$	moyen	oui	excellent	oui .
$> 40$	faible	oui	excellent	non
$> 40$	faible	oui	bon	oui .
$> 40$	moyen	oui	bon	oui .
31-40	faible	oui	excellent	oui .
31-40	eleve	oui	bon	oui .

# Principe

- Diviser l'ensemble des données d'apprentissage successivement en sous-groupes, selon les valeurs prises par les variables explicatives qui, à chaque étape, discrimine le mieux la variable cible.
- On commence par choisir la variable qui sépare le mieux les observations de façon à avoir des sous-groupes, qu'on appelle nœuds, les plus homogènes possibles par rapport à la variable cible.
- On réitère le processus pour chaque nœud *fils* jusqu'à ce qu'il ne soit plus possible de continuer (ex: noeuds purs) ou on a atteint un certain critère d'arrêt.

Le modèle qui en résulte est une longue série de règles simples.

# Vocabulaire

- Racine
- Noeud
- Noeud terminaux (feuilles)
- Taux d'erreur dans un noeud
- Taux d'erreur global
- Hauteur de l'arbre

Profondeur

## Deux questions

- 1 Comment choisir la bonne variable pour séparer?
- 2 Quand arrêter?

# Principaux algorithmes

- CHAID

Chi-Square Automatic Interaction Detection Kass, GV (1980).  
An Exploratory Technique for Investigating Large Quantities of  
Categorical Data. Applied Statistics, vol 29, no 2, pp 119-127.

- CART

Classification And Regression Trees (Breiman et al., 1984)



## CHAID

# Objectif

- Comprendre comment construire un arbre de type CHAID.

## CHAID (approche conditionnelle)

- Variables cible et explicatives nominales ou ordinales
- Les embranchements sont binaires ou non (facilite l'interprétation)

## CHAID (approche conditionnelle)

### Comment choisir les variables à chaque nœud?

#### ■ Test du $\chi^2$ <sup>1</sup>

	p-valeur
Age	0.33
Revenu	1.00
Etudiant	0.26
Credit	0.56

---

<sup>1</sup>Ici obtenu par simulation parce que l'échantillon est trop petit

## En pratique

La fonction `ctree` de la librairie `party` permet de construire des arbres de type CHAID.<sup>2</sup>

```
library(party)
controle2 <- ctree_control(mincriterion = 0.6,
                           minsplit = 2,
                           minbucket = 1)

arbre.ctree <- ctree(achat ~ .,
                     don,
                     control = controle2)
```

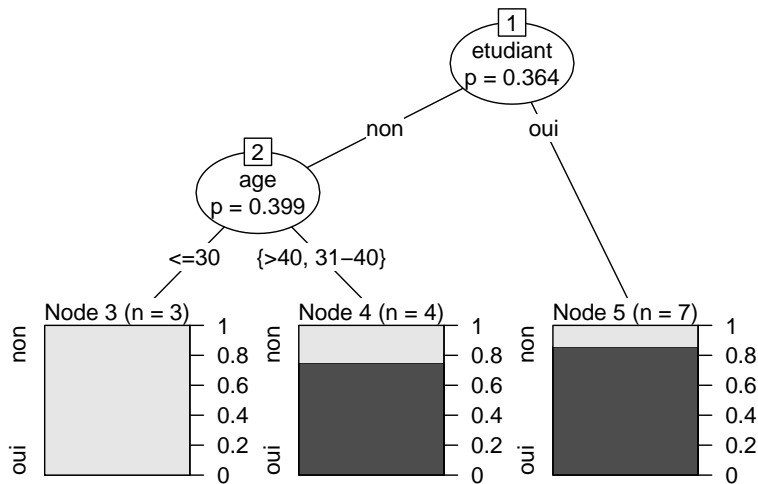
---

<sup>2</sup>Lien vers la section sur les arbres dans les notes de cours.

## Exemple

```
##
##   Conditional inference tree with 3 terminal nodes
##
## Response:  achat
## Inputs:  age, revenu, etudiant, credit
## Number of observations:  14
##
## 1) etudiant == {non}; criterion = 0.636, statistic = 3.293
##   2) age == {<=30}; criterion = 0.601, statistic = 4.25
##     3)* weights = 3
##     2) age == {>40, 31-40}
##       4)* weights = 4
##     1) etudiant == {oui}
##       5)* weights = 7
```

# Exemple



## Critère d'arrêt

### Quand arrêter?

- Lorsque  $p\text{-valeur} > \text{seuil}$
- Effectif minimal dans les noeuds terminaux
- Hauteur maximale



CART<sup>3</sup>

---

<sup>3</sup>Lire la section 9.2 de ESL

# Objectif

- Comprendre comment se construit un arbre de type CART

# CART

## Comment choisir les variables à chaque noeud?

Considère tous les embranchements *binaires* possibles pour toutes les variables explicatives<sup>4</sup>

On sélectionne généralement sur la base de

- EQM
- Indice de Gini

---

<sup>4</sup>Avec  $m$  valeurs distinctes,  $2m-1$  regroupements binaires possibles pour les variables nominales;  $m-1$  points milieux distincts pour les variables continues. . . c'est long!

# CART

Indice de Gini (indice d'impureté)

$$Gini = \sum_{k=1}^K p_k(1 - p_k) = 1 - \sum_{k=1}^K p_k^2$$

# CART

## Quand arrêter?

- Effectif minimal dans les noeuds terminaux
- Hauteur maximale

En général, on doit faire de l'élagage sur la base de la complexité.

# CART

Coût de complexité d'un arbre  $T$  (équation 9.16).

$$C_{\alpha}(T) = \sum_{w=1}^{|T|} N_w Q_w + \alpha |T|$$

où

- $|T|$  est le nombre de noeuds terminaux dans l'arbre  $T$
- $Q_w$  est l'impureté ou l'erreur quadratique moyenne ( $E[(y_i - \hat{y}_i)^2]$ ) dans le noeud  $w$
- $\alpha$  est le paramètre de pénalisation

## En pratique

Vous pouvez utiliser la librairie `rpart` pour créer ce type d'arbre.<sup>5</sup>

```
library(rpart)
library(rpart.plot)

# Hyperparamètres
hp <- list(split = 'gini')
controles_rpart <- rpart.control(minsplit = 1)

# Construction de l'arbre
arbre_rpart <- rpart(achat~.,don,
                      parms = hp,
                      control = controles_rpart)
```

---

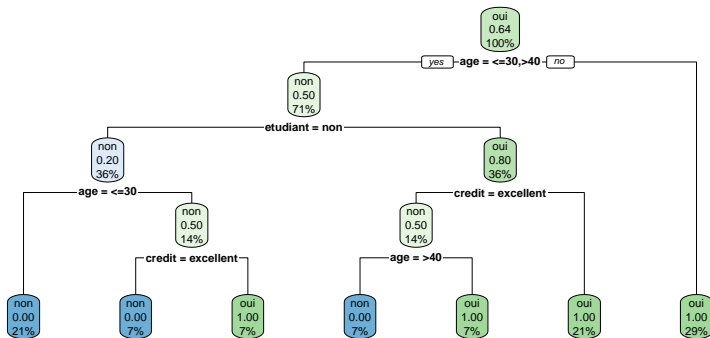
<sup>5</sup>Lien vers la section sur les arbres dans les notes de cours.

# Exemple

```
## n= 14
##
## node), split, n, loss, yval, (yprob)
##      * denotes terminal node
##
## 1) root 14 5 oui (0.3571429 0.6428571)
##    2) age=<=30,>40 10 5 non (0.5000000 0.5000000)
##      4) etudiant=non 5 1 non (0.8000000 0.2000000)
##        8) age=<=30 3 0 non (1.0000000 0.0000000) *
##        9) age=>40 2 1 non (0.5000000 0.5000000)
##          18) credit=excellent 1 0 non (1.0000000 0.0000000) *
##          19) credit=bon 1 0 oui (0.0000000 1.0000000) *
##      5) etudiant=oui 5 1 oui (0.2000000 0.8000000)
##        10) credit=excellent 2 1 non (0.5000000 0.5000000)
##          20) age=>40 1 0 non (1.0000000 0.0000000) *
##          21) age=<=30 1 0 oui (0.0000000 1.0000000) *
##      11) credit=bon 3 0 oui (0.0000000 1.0000000) *
##    3) age=31-40 4 0 oui (0.0000000 1.0000000) *
```



## Exemple



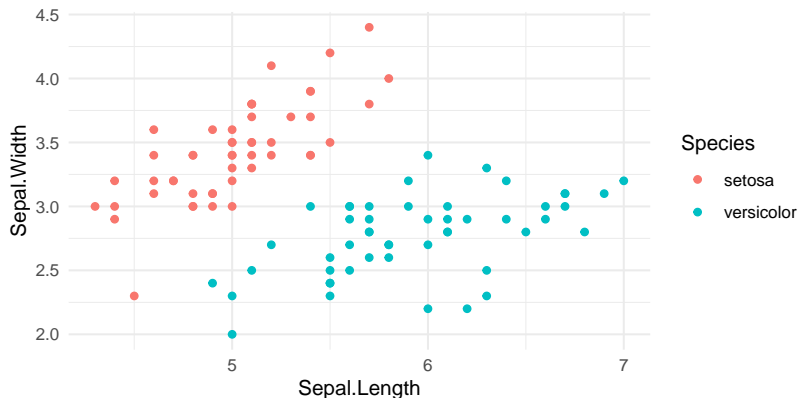
Lien vers la visualisation interactive

# Mathématique

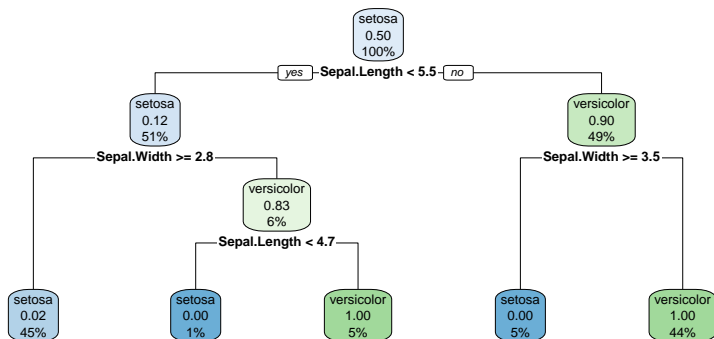
# Objectif

- Comprendre la formulation mathématique des arbres.

On peut voir les arbres comme un partitionnement de l'espace en  $M$  régions  $R_1, R_2, \dots, R_M$ .



Le modèle s'écrit  $f(\mathbf{x}) = \sum_{m=1}^M c_m \mathbf{I}(x \in R_m)$  (eq. 9.10).



## Discussion

# Objectif

- Comprendre les avantages et inconvénients des arbres
- Comprendre les alternatives possibles

# Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données



## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes

## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter

## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter
- 4 Tient implicitement compte des interactions possibles entre les variables

## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter
- 4 Tient implicitement compte des interactions possibles entre les variables
- 5 Sélectionne implicitement les variables importantes

# Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter
- 4 Tient implicitement compte des interactions possibles entre les variables
- 5 Sélectionne implicitement les variables importantes
- 6 Permet d'obtenir un modèle non linéaire

## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter
- 4 Tient implicitement compte des interactions possibles entre les variables
- 5 Sélectionne implicitement les variables importantes
- 6 Permet d'obtenir un modèle non linéaire
- 7 Invariante aux transformation monotones des prédicteurs

## Avantages

- 1 Aucune hypothèse *a priori* n'est faite sur la distribution des données
- 2 Robuste aux données extrêmes
- 3 Facile à interpréter
- 4 Tient implicitement compte des interactions possibles entre les variables
- 5 Sélectionne implicitement les variables importantes
- 6 Permet d'obtenir un modèle non linéaire
- 7 Invariante aux transformation monotones des prédicteurs
- 8 Permet une gestion intéressante des valeurs manquantes.

## Traitement des valeurs manquantes

- Le choix de la variable et de la meilleure séparation se fait, à chaque noeud, uniquement avec les données non manquantes.
- On utilise ensuite des variables substitut pour prédire à quel noeud appartient chaque observation.



# Inconvénients

- 1 L'approche n'est pas particulièrement performante en termes de prédiction

## Inconvénients

- 1 L'approche n'est pas particulièrement performante en termes de prédiction
- 2 Les risque de surapprentissage sont élevés

# Inconvénients

- 1 L'approche n'est pas particulièrement performante en termes de prédiction
- 2 Les risque de surapprentissage sont élevés
- 3 Nous ne sommes pas certains d'atteindre la meilleure partition

# Inconvénients

- 1 L'approche n'est pas particulièrement performante en termes de prédiction
- 2 Les risque de surapprentissage sont élevés
- 3 Nous ne sommes pas certains d'atteindre la meilleure partition
- 4 Peut être **instable** dans les résultats qu'elle produit

## Comparaison CART et CHAID

- Performance prédictive similaire
- L'approche de type cart tendent à choisir les variables qui ont le plus de modalité (et les variables continues). Les approches conditionnelles sont moins sensibles à ce problème.

## Adaptation à des problèmes particuliers

On peut facilement modifier le critère de séparation pour tenir compte de situations particulières.

Ex. Les forêts de survie