

Séance 10: Arbres de classifications et Forêts aléatoires

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Dixième partie X

CART : Classification And Regression Trees

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Introduction

- Méthode pour **modéliser une classification ou une régression**
- Apporte des solutions graphiques facilement interprétables
- Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives
- Est capable de gérer à la fois les variables quantitatives ET qualitatives
- Acrônymes courants : CART/C4.5
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les **variables sont nombreuses**
- Algorithme récursif donc un petit peu calculatoire

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de 2 sous-arbres
- Un arbre de profondeur 0 est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - ① La définition d'un critère permettant de sélectionner la meilleure division possible parmi toutes celles d'admissibles
 - ② Une règle permettant de décider qu'un noeud est terminal
 - ③ L'affectation à chaque feuille une classe (classification) ou valeurs réelles (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de 2 sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - ① La définition d'un critère permettant de sélectionner la meilleure division possible parmi toutes celles d'admissibles
 - ② Une règle permettant de décider qu'un noeud est terminal
 - ③ L'affectation à chaque feuille une classe (classification) ou valeurs réelles (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - Une règle permettant de **décider qu'un noeud est terminal**
 - L'affectation à **chaque feuille une classe (classification) ou valeurs réelles (régression)**
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - Une règle permettant de **décider qu'un noeud est terminal**
 - L'affectation à chaque feuille une **classe (classification) ou valeurs réelles (régression)**
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - 1 La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - 2 Une règle permettant de **décider qu'un noeud est terminal**
 - 3 **L'affectation à chaque feuille une classe** (classification) ou **valeurs réelles** (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - 1 La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - 2 Une règle permettant de **décider qu'un noeud est terminal**
 - 3 L'affectation à **chaque feuille une classe** (classification) ou **valeurs réelles** (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - 1 La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - 2 Une règle permettant de **décider qu'un noeud est terminal**
 - 3 L'affectation à **chaque feuille une classe** (classification) ou **valeurs réelles** (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - 1 La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - 2 Une règle permettant de **décider qu'un noeud est terminal**
 - 3 **L'affectation à chaque feuille une classe** (classification) ou **valeurs réelles** (régression)
- C'est le second point qui est le plus délicat

Algorithme de construction

- Variables explicatives $X^1, \dots, X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n -échantillon, qualitative à m modalités ou quantitative
- Un arbre est défini récursivement par la donnée d'un noeud et de **2** sous-arbres
- Un **arbre de profondeur 0** est considéré comme une feuille et contient une règle de décision
- L'algorithme nécessite donc :
 - 1 La définition d'un critère permettant de **sélectionner la meilleure division possible** parmi toutes celles d'admissibles
 - 2 Une règle permettant de **décider qu'un noeud est terminal**
 - 3 **L'affectation à chaque feuille une classe** (classification) ou **valeurs réelles** (régression)
- C'est le second point qui est le plus délicat

Exemple d'un arbre de classification

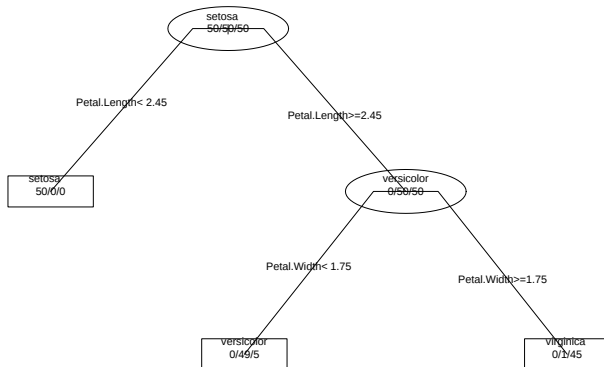


FIG.: Un exemple célèbre d'application de CART en classification sur 4 variables pour 2 classes

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- **La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène**

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- **La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène**

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- La croissance de l'arbre s'arrête dès que le noeud obtenu est **homogène**

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène

Critère de division

- Une division est **admissible** si aucun des noeuds qui en découle est vide.
- Le critère de division repose sur la définition d'une fonction d'**hétérogénéité**
- Cette fonction doit satisfaire les 2 propriétés :
 - Elle doit être **nulle si et seulement si le noeud qui en découle est homogène** : tous les individus appartiennent à la même classe
 - Elle doit être **maximale lorsque les valeurs de Y sont équiprobables** (très dispersées)
- La division de chaque noeud génère un fils gauche et un fils droit et la division retenue sera celle qui minimise la somme des désordres des deux fils obtenus.
- Graphiquement : la longueur d'une branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division
- **La croissance de l'arbre s'arrête dès que le noeud obtenu est homogène**

Critère de décision

- Pour une variable qualitative : on renvoie la classe de tous les individus associés à chacun des noeuds terminaux, ou bien celle qui est majoritaire
- Pour une variable quantitative : on renvoie la moyenne de tous les individus associés à chacun des noeuds terminaux.

Critère de décision

- Pour une variable qualitative : on renvoie **la classe de tous les individus associés à chacun des noeuds terminaux**, ou bien celle qui est **majoritaire**
- Pour une variable quantitative : on renvoie la **moyenne de tous les individus** associés à chacun des noeuds terminaux.

Critère d'homogénéité

Là encore, il y a deux cas :

- Y est **quantitative** et on peut faire une partition des individus en J classes.
- Chaque élément de la partition en J sous-arbres contient n_j individus ($j = 1 \dots J$)
- On mesure l'hétérogénéité de la partition *via*

$$D = \sum_{j=1}^J D_j = \sum_{j=1}^J (\mu_{i,j} - \mu_{.,j})^2 \quad \text{où} \quad \mu_{.,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{i,j}$$

- La différence d'hétérogénéité vaut donc :

$$\Delta = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,.})^2 - \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,j})^2$$

- On peut établir la formule qui rend minimale la variance intra-classe

$$\Delta = \sum_{j=1}^J n_j (\mu_{.,.} - \mu_{.,j})^2$$

Critère d'homogénéité

Là encore, il y a deux cas :

- Y est **quantitative** et on peut faire une partition des individus en J classes.
- Chaque élément de la partition en J sous-arbres contient n_j individus ($j = 1 \dots J$)
- On mesure l'hétérogénéité de la partition *via*

$$D = \sum_{j=1}^J D_j = \sum_{j=1}^J (\mu_{i,j} - \mu_{.,j})^2 \quad \text{où} \quad \mu_{.,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{i,j}$$

- La différence d'hétérogénéité vaut donc :

$$\Delta = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,.})^2 - \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,j})^2$$

- On peut établir la formule qui rend minimale la variance intra-classe

$$\Delta = \sum_{j=1}^J n_j (\mu_{.,.} - \mu_{.,j})^2$$

Critère d'homogénéité

Là encore, il y a deux cas :

- Y est **quantitative** et on peut faire une partition des individus en J classes.
- Chaque élément de la partition en J sous-arbres contient n_j individus ($j = 1 \dots J$)
- On mesure l'hétérogénéité de la partition *via*

$$D = \sum_{j=1}^J D_j = \sum_{j=1}^J (\mu_{i,j} - \mu_{.,j})^2 \quad \text{où} \quad \mu_{.,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{i,j}$$

- La différence d'hétérogénéité vaut donc :

$$\Delta = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,.})^2 - \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,j})^2$$

- On peut établir la formule qui rend minimale la variance intra-classe

$$\Delta = \sum_{j=1}^J n_j (\mu_{.,.} - \mu_{.,j})^2$$

Critère d'homogénéité

Là encore, il y a deux cas :

- Y est **quantitative** et on peut faire une partition des individus en J classes.
- Chaque élément de la partition en J sous-arbres contient n_j individus ($j = 1 \dots J$)
- On mesure l'hétérogénéité de la partition *via*

$$D = \sum_{j=1}^J D_j = \sum_{j=1}^J (\mu_{i,j} - \mu_{.,j})^2 \quad \text{où} \quad \mu_{.,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{i,j}$$

- La différence d'hétérogénéité vaut donc :

$$\Delta = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,.})^2 - \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,j})^2$$

- On peut établir la formule qui rend minimale la variance intra-classe

$$\Delta = \sum_{j=1}^J n_j (\mu_{.,.} - \mu_{.,j})^2$$

Critère d'homogénéité

Là encore, il y a deux cas :

- Y est **quantitative** et on peut faire une partition des individus en J classes.
- Chaque élément de la partition en J sous-arbres contient n_j individus ($j = 1 \dots J$)
- On mesure l'hétérogénéité de la partition *via*

$$D = \sum_{j=1}^J D_j = \sum_{j=1}^J (\mu_{i,j} - \mu_{.,j})^2 \quad \text{où} \quad \mu_{.,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{i,j}$$

- La différence d'hétérogénéité vaut donc :

$$\Delta = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,.})^2 - \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{i,j} - \mu_{.,j})^2$$

- On peut établir la formule qui rend minimale la variance intra-classe

$$\Delta = \sum_{j=1}^J n_j (\mu_{.,.} - \mu_{.,j})^2$$

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Critère d'homogénéité

- Y est **qualitative** et l'hétérogénéité est définie à partir de la notion d'entropie, ou du critère de concentration de Gini
- En général, le choix du critère est moins important que l'élagage de l'arbre de classification
- Y à m modalités : l'arbre engendre une partition des individus.
- On considère la k -ième classe. On peut alors former :

$$p_{l,k} = P(Y = C_l | k)$$

- On mesure l'hétérogénéité expliquée par la classe k :

$$D_k = -2 \sum_{l=1}^m n_{.,k} p_{l,k} \log(p_{l,k})$$

- Le désordre total expliqué est alors $D = \sum_{k=1}^K D_k$
- Chaque D_k est positif, c'est la fonction d'entropie de $p_{.,k}$
- D_k est nul si la probabilité $p_{.,k}$ est une masse de Dirac, elle est maximale si $p_{.,k}$ est uniforme.

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

- Dans des situations complexes, les arbres construits peuvent être extrêmement raffinés
- Cela rend la nature de CART **instable** car fortement dépendant des échantillons qui ont permis leur estimation
- Il est donc nécessaire, d'un point de vue statistique, d'élaguer l'arbre (*pruning*)
- Le principe consiste à construire l'arbre maximal ainsi qu'une suite de sous-arbres emboîtés
- On mesure la qualité de discrimination de l'arbre par

$$D = \sum_{k=1}^K D_k(A)$$

où D_k désigne le nombre de mal-classés, où la variance sur la feuille k

- On pénalise cette qualité par

$$C(A) = D(A) + \gamma K$$

Elagage

Procédé d'élagage :

- Si $\gamma = 0$, on obtient l'arbre maximal
- On augmente progressivement γ jusqu'à ce que l'une des divisions de A_K (celle pour laquelle D est la plus faible) apparaît comme superflue.
- On obtient alors A_{K-1} en fusionnant les feuilles terminales (élagage 1)
- On itère le procédé.
- On choisit l'arbre optimal par validation croisée comme celui qui minimise la déviance.

Elagage

Procédé d'élagage :

- Si $\gamma = 0$, on obtient l'arbre maximal
- On augmente progressivement γ jusqu'à ce que l'une des divisions de A_K (celle pour laquelle D est la plus faible) apparaît comme superflue.
- On obtient alors A_{K-1} en fusionnant les feuilles terminales (élagage 1)
- On itère le procédé.
- On choisit l'arbre optimal par validation croisée comme celui qui minimise la déviance.

Elagage

Procédé d'élagage :

- Si $\gamma = 0$, on obtient l'arbre maximal
- On augmente progressivement γ jusqu'à ce que l'une des divisions de A_K (celle pour laquelle D est la plus faible) apparaît comme superflue.
- On obtient alors A_{K-1} en fusionnant les feuilles terminales (élagage 1)
- On itère le procédé.
- On choisit l'arbre optimal par validation croisée comme celui qui minimise la déviance.

Elagage

Procédé d'élagage :

- Si $\gamma = 0$, on obtient l'arbre maximal
- On augmente progressivement γ jusqu'à ce que l'une des divisions de A_K (celle pour laquelle D est la plus faible) apparaît comme superflue.
- On obtient alors A_{K-1} en fusionnant les feuilles terminales (élagage 1)
- On itère le procédé.
- On choisit l'arbre optimal par validation croisée comme celui qui minimise la déviance.

Elagage

Procédé d'élagage :

- Si $\gamma = 0$, on obtient l'arbre maximal
- On augmente progressivement γ jusqu'à ce que l'une des divisions de A_K (celle pour laquelle D est la plus faible) apparaisse comme superflue.
- On obtient alors A_{K-1} en fusionnant les feuilles terminales (élagage 1)
- On itère le procédé.
- On choisit l'arbre optimal par validation croisée comme celui qui minimise la déviance.

Elagage

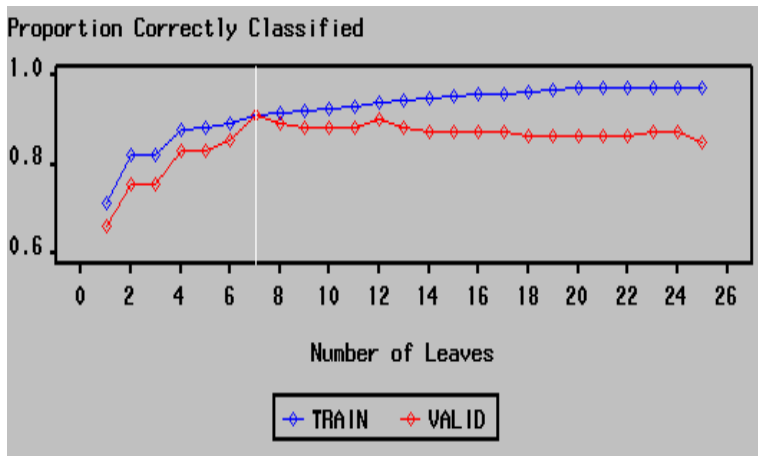
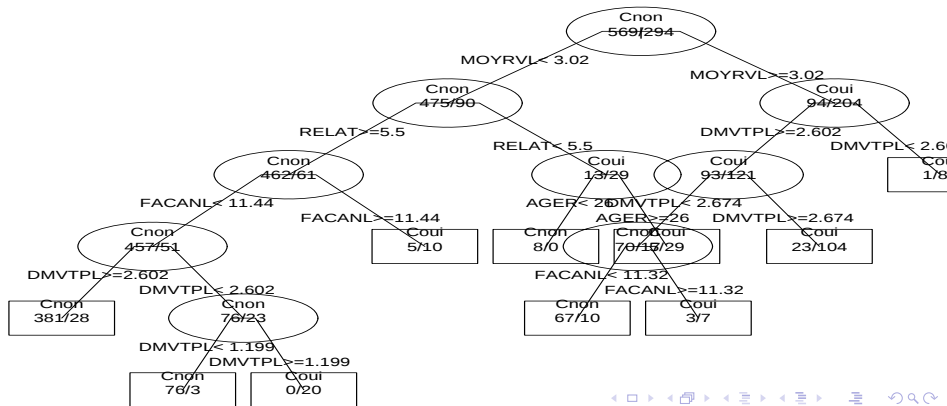


FIG.: Banque : choix du nombre de feuilles par échantillon de validation (SEM, 2001).

Elagage

La librairie `rpart` de R propose d'optimiser l'élagage par validation croisée. L'arbre ainsi obtenu sur Visa premier :

Endpoint = CARVP



Onzième partie XI

Random Forests : Agrégation d'arbres de classifications pour la construction de forêts aléatoires