Séance 5: Arbres de classifications

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Cinquième partie V

CART : Classification And Regression Trees

Introduction - Cadre d'application

- Méthode pour modéliser une classification ou une régression
- C'est un algorithme "décisionnel" construit à partir d'un training set et à tester sur un test set
- Acrônymes courants : CART/C4.5

Points forts :

- Apporte des solutions graphiques facilement interprétables
- Est capable de gérer à la fois les variables quantitatives ET qualitatives simultanément
- Gère des tailles d'échantillons importantes
- Méthode adaptée au cas où les variables sont nombreuses

Attention : Algorithme récursif donc un petit peu calculatoire

Cadre du modèle

- Variables explicatives $X^1, \ldots X^p$ (qualitatives ou quantitatives)
- Variable Y à expliquer observée sur un n-échantillon, qualitative à m modalités ou quantitative.

Définition : Un arbre est défini récursivement par la donnée d'un noeud et de 2 sous-arbres.

Exemple : Un arbre de profondeur 0 est considéré comme une feuille et correspond à une règle de décision unique.

Remarque : Plus généralement, chaque nœud de l'arbre contient une règle de décision pour la donnée à prédire. On parcourt l'arbre de façon descendante en appliquant au fur et à mesure les tests logiques sur les signaux.
Idée : découper *via* des hyperplans l'espace engendré par des variables explicatives

Algorithme de construction

- 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 nœud est terminal
  3. L'affectation à chaque feuille une classe (classification) ou valeurs réelles (régression)
- La stratégie du premier point dépend de l'objectif : classification/régression.
- C'est le second point qui est le plus délicat.

Exemple d'un arbre de classification

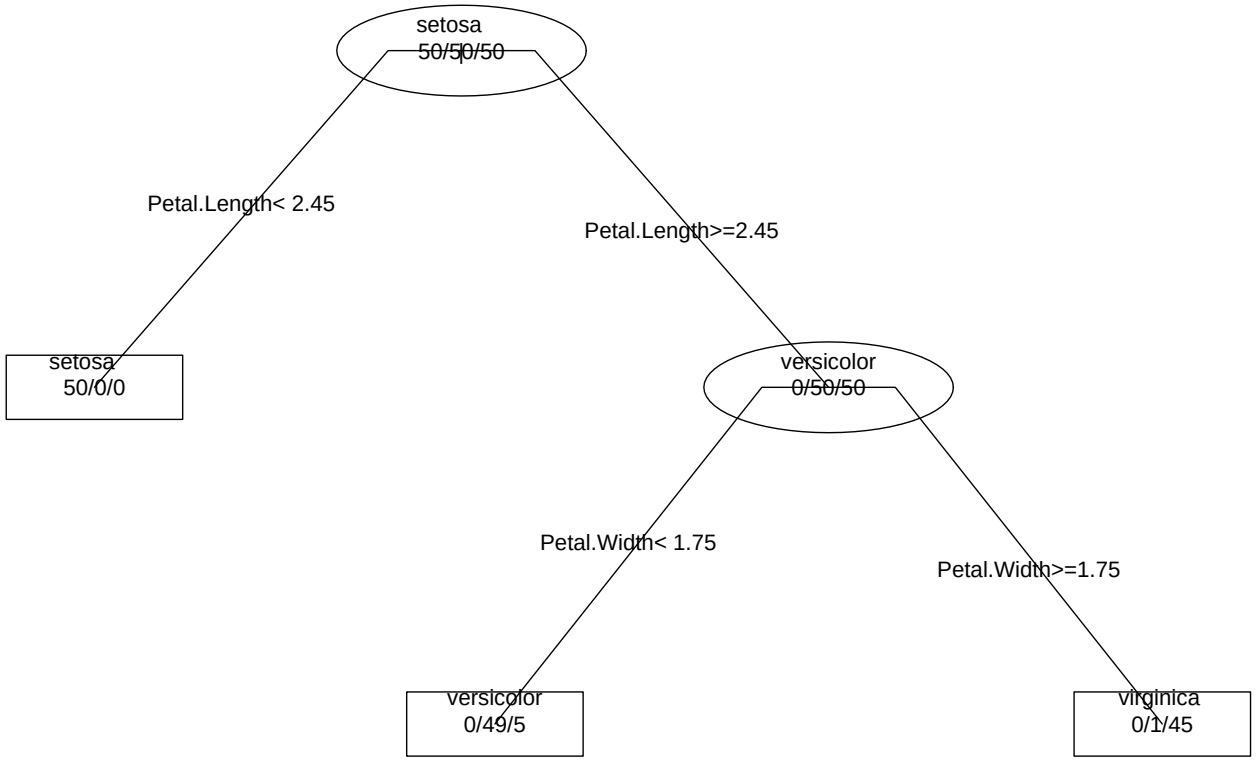

FIG : Un exemple célébrissime d'application de CART en classification sur 4 variables pour 3 classes

Critère de division

Impératif pour la 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 pour la classification comme pour la régression :
  - 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)

Critère de division

Résultat d'une division :

- La division de chaque nœud 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 nœud obtenu est homogène

Critère de décision

Décision pour les nœuds terminaux :

- 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 - Variables Quantitatives

L'objectif est donc de faire une régression pour prédire les valeurs de Y.

- 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*

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

Critère d'homogénéité - Variables Quantitatives

- Si l'on considère aucune partition, l'hétérogénéité vaut :

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

- La différence d'hétérogénéité entre "sans partition" et "avec" vaut donc :

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

- Théorème : Le différentiel d'hétérogénéité vaut :

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

Critère d'homogénéité - Variables Quantitatives

Dans le cas où l'on effectue une partition binaire, on a $J = 2$. On obtient donc

$$\Delta = n_1 n_2 (\mu_{.,1} - \mu_{.,2})^2$$

Pour la construction récursive de l'arbre de régression, il s'agit de maximiser Δ (on rend l'arbre le plus homogène possible).

Algorithme :

Étape 1 : Choisir le seuil tel que Δ est maximal (dépend du calcul des différentes moyennes des deux sous-groupes d'individus).

Étape n :

1. On parcourt tous les nœuds possibles de l'arbre à l'étape n et on énumère pour chacun les meilleurs "splits"
2. On retient le meilleur parmi tous les nœuds
3. Si le nouvel arbre n'est pas homogène, on fait l'étape $n + 1$

Critère d'homogénéité - Variables Qualitatives

Le cadre est donc la classification d'une variable Y qualitative.

- 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 a m modalités : l'arbre engendre une partition des individus.
- On considère la k-ième classe. On peut alors estimer empiriquement :

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

C'est la probabilité pour un individu de la partition k de l'arbre d'avoir $Y = C_l$.

Critère d'homogénéité - Variables Qualitatives

- Définition : 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 $D = \sum_{k=1}^{K} D_k$
- Proposition : 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 - Variables Qualitatives

Algorithme : C'est très comparable à celui qui concerne les variables quantitatives

Étape 1 : On choisit la règle de décision qui maximise l'homogénéité des classes, autrement dit l'hétérogénéité doit être minimale

Étape n : On parcourt tous les nœuds terminaux de l'arbre et on retient la meilleure partition (celle qui maximise l'homogénéité).

Étape n : On itère tant que l'homogénéité est "améliorable".

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

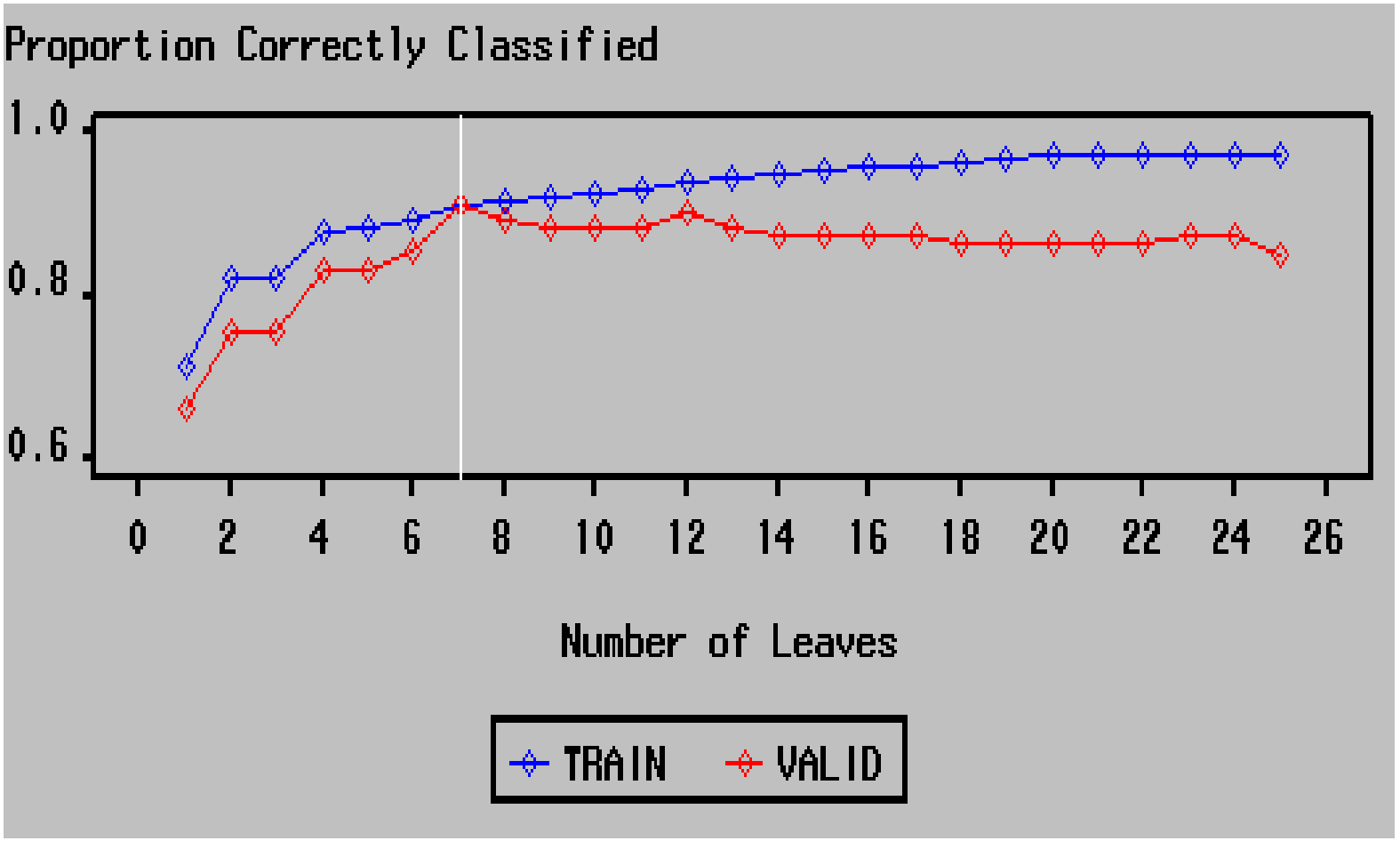

FIG : Banque : choix du nombre de feuilles par échantillon de validation (SEM, 2001)

Elagage

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