Chapitre 6

Arbres binaires de décision

6.1 Introduction

Les arbres de décision sont des outils de modélisation très utilisés pour la classification supervisée (prédiction d'une variable nominale) ou la régression. Leur popularité est due au fait qu'on les représente graphiquement sous la forme d'un arbre simple à interpréter, même pour les néophytes. Ils constituent ainsi une aide efficace pour l'aide à la décision.

Ces méthodes sont efficaces pour des tailles d'échantillons importantes mais elles requièrent plutôt moins d'hypothèses que d'autres méthodes statistiques classiques. Les arbres de décision sont particulièrement bien adaptés aux situations où les variables explicatives sont nombreuses car la procédure de sélection de variables est intégrée à l'algortihme construisant l'arbre.
Exemple : risque pour un crédit (voir figure ??).

Les algorithmes les plus répandus pour la construction de arbres binaires sont l'algorithme CART (Breiman, 1984) et l'algorithme C4.5 (Quinlan, 1993).
Référence : cours de Ph. Besse.

6.2 Construction d'un arbre binaire

6.2.1 Principe

Soit un échantillon de taille n constitué de l'observation de p variables quantitatives ou qualitatives $X_1, \cdots, X_p$ et d'une variable à expliquer qualitative Y à m modalités.

La construction d'un arbre de discrimination binaire consiste à déterminer une séquence de noeuds.

- Un noeud est défini par le choix conjoint d'une variable parmi les explicatives et d'une division qui induit une partition en deux classes. Implicitement, à chaque noeud correspond donc un sous-ensemble de l'échantillon auquel est appliquée une dichotomie.

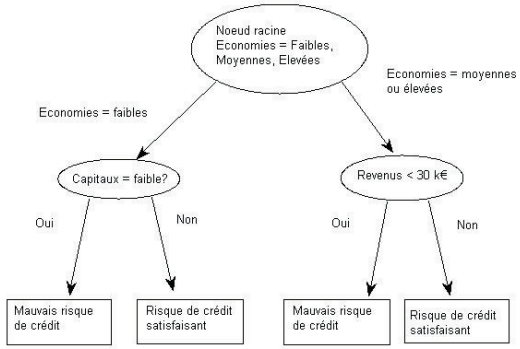

FIGURE 6.1 – Exemple d'arbre de décision binaire

- Une division est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.
- A la racine ou noeud initial correspond l'ensemble des individus de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles.

L'algorithme nécessite donc

1. La définition d'un critère permettant de sélectionner la "meilleure" division parmi toutes celles admissibles pour les différentes variables ;
2. une règle permettant de décider qu'un noeud est terminal : il devient ainsi une feuille ;
3. l'affectation de chaque feuille à l'une des classes ou à une valeur à expliquer.

Le point 2 est le plus délicat. Il correspond à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à surparamétrisation est instable et donc souvent défaillllant pour la prévision. La contribution majeure de Breiman et al. (1984) est justement une stratégie de recherche de l'arbre optimal. Elle consiste à

1. construire l'arbre maximal,
2. ordonner les sous-arbres selon une séquence emboitée suivant la décroissance d'un critère de déviance ou de taux de mal classés,
3. puis à sélectionner le sous-arbre optimal ; c'est la procédure d'élagage.

6.2.2 Critère de division

Une division est dite admissible si aucun des deux noeuds descendant qui en découle n'est vide.

Exemples -

- Cas d'une variable ordinale à m modalités : on obtient $(m-1)$ divisions admissaibles
- Cas d'une variable nominale à m modalités : on obtient $2^(m-1)-1$ divisions admissibles.
- Cas d'une variable qualitative : on cherche des seuils de division qui séparent le domaine de définition de la variable en 2 intervalles.

Le critère de division repose sur une notion d'hétérogénéité. L'objectif est de partager les individus en deux groupes les plus homogènes possibles au sens de la variable à expliquer. L'hétérogénéité se mesure par une fonction non négative qui doit être

- nulle si et seulement si le noeud est homogène ;
- maximale lorsque les valeurs de Y sont ésuiprobables.

La division du noeud créé deux fils. Parmi toutes les divisions admissibles du noeud k, l'algorithme retient celle qui rend la somme $D_{(k+1)} + D_{(k+2)}$ des désordres des noeuds fils minimale.

6.2.3 Règle d'arrêt

La croissance de l'arbre s'arrête à un noeud donnée, qui devient donc terminal, lorsqu'il est homogène ou si le nombre d'observations qu'il contient est inférieur à une valeur seuil choisie (en général entre 1 et 6).

6.2.4 Affectation

. Dans le cas où Y est qualitative, chaque feuille est affectée à une classe de Y en considérant :

- la classe la mieux représentée dans le noeud ;
- la classe a posteriori la plus probable au sens bayésien si des probabilité a priori sont connues ;
- la classe la moins coûteuse si des coûts de mauvais classement sont donnés.

Dans le cas où Y est quantitative, on affecte généralement la moyenne des observations associées à cette feuille

6.3 Critères homogénéité

Nous considérons ici le cas où la variable à expliquer Y est qualitative. Dans ce cas la fonction d'hétérogénéité est définie à partir de la notion d'entropie, du critère de Gini ou encore un d'une statistique du χ^2. En pratique on vérifie que le choix du critère importe moins que celui du niveau d'élagage.

6.3.1 Entropie

On note les modalités C_l, $l = 1, \cdots, m$ les m modalités de la variable Y. L'arbre induit une partition pour laquelle n_k désigne l'effectif du $k$ème noeud. Soit

$$p_{lk} = P[C_l|k] \text{ avec } \sum_{l=1}^{m} p_{lk} = 1$$

la probabilité qu'un élément du $k$ème noeud appartienne à la $l$ième classe.

Le désordre du $k$ème noeud, défini à partir de l'entropie, s'écrit avec la convention $0 \log(0) = 0$:

$$D_k = -2 \sum_{l=1}^{m} n_k p_{lk} \log(p_{lk})$$

Tandis que l'hétérogénéité de la partition est donnée par :

$$D = \sum_{k=1}^{K} D_k = -2 \sum_{k=1}^{K} \sum_{l=1}^{m} n_k p_{lk} \log(p_{lk})$$

Remarques ;

- La quantité D s'apparente à une déviance (pour une variable mutinomiale).
- Cette quantité est positive ou nulle et elle est nulle seulement si les probabilités p_{lk} ne prennent que des valeurs nulles sauf une égale à 1 correspondant à l'absence de mélange.

Ces quantités sont estimées en utilisant les probabilités empiriques c'est à dire des comptages :

$$\hat{D}_k = -2\sum_{l=1}^{m} n_k \frac{n_{lk}}{n_k} \log(\frac{n_{lk}}{n_k})$$

et

$$\hat{D} = \sum_{k=1}^{K} \hat{D}_k$$

Exemple - Considérons l'arbre de la figure 6.2.

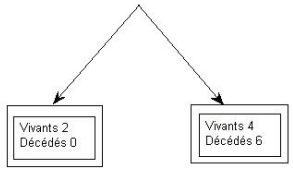

FIGURE 6.2 – Exemple d'arbre de décision binaire

$$\hat{D}_1 = -2\left(2\frac{2}{2}\log(\frac{2}{2}) + 2\frac{0}{2}\log(\frac{0}{2})\right) = 0$$
$$\hat{D}_2 = -2\left(10\frac{4}{10}\log(\frac{4}{10}) + 10\frac{6}{10}\log(\frac{6}{10})\right) = 13.5$$
$$\hat{D} = 13.5$$

Pour ce modèle la logvraisemblance

$$\log L = \text{Cste} + \sum_{k=1}^{K}\sum_{l=1}^{m} n_{lk}\log(p_{lk})$$

est rendue maximale pour

$$\mathcal{L}_\mu = \sup_{p_{lk}} \log L = \text{Cste} + \sum_{k=1}^{K}\sum_{l=1}^{m} n_{lk}\log(\frac{n_{lk}}{n_k})$$

Pour le modèle saturé (une catégorie par objet), cet optimum prend la valeur de la constante et la déviance par rapport au modèle saturé s'exprime comme :

$$\mathcal{D} = -2\sum_{k=1}^{K}\sum_{l=1}^{m} n_{lk}\log(\frac{n_{lk}}{n_k}) = \hat{D}$$

6.3.2 Indice de Gini

Le critère basé sur l'entropie peut être remplacé par l'indice de Gini $1 - \sum_{l=1}^{m} p_{lk}^2$ qui conduit à une autre définition de l'hétérogénéité également utilisée mais qui ne s'interprète plus en terme de déviance.

6.3.3 Distance du χ^2

Pour évaluer la pertinence de la variable dans la segmentation, l'algortihme CHAID propose d'utiliser le critère du chi 2 qui mesure un écart à l'indépendance. Nous utilisons ici les notations suivantes

Y (cible) / X	x_1	x_k	x_K	$\sum$
y_1				
		$\cdots$		
y_l	$\cdots$	n_{lk}	$\cdots$	$n_{.k}$
		$\cdots$		
y_K				
$\sum$		$n_{l.}$		n

Et la distance du chi 2 s'exprime alors comme suit :

$$\chi^2 = \sum l = 1^m \sum k = 1^K \frac{n_{lk} - \frac{n_l . n.k}{n}}{\frac{n_l . n.k}{n}}$$

Ce critère varie de 0 à $+\infty$. Il n'est pas facile à manipuler car il avantage les variables ayant un nombre élévé de modalités. On peut le normaliser par le nombre de dégrés de libertés en se ramenant au T de Tschuprow.

6.4 Élagage

Dans les situations complexes, le procédé de la construction d'arbre conduit à des arbres très rafinés et donc à des modèles de prévision très instables. Ils sont trop dépendants des échantillons qui ont permis l'estimation. On se trouve dans une situation de sur-ajustement et on est donc amené à rechercher un modèle plus parcimonieux.

Une méthode consiste à construire une suite de sous arbres emboités puis à procéder à l'élagage de l'arbre maximal. On choisi alors de retenir l'arbre optimal au sens d'un critère (par exemple le taux d'individus mal classés).

6.4.1 Construction de la séquence d'arbres

Pour un arbre donné A on note K le nombre de feuilles ou noeuds terminaux de A ; la valeur de k exprime la complexité de l'arbre. La mesure de la qualité de discrimination s'exprime par un critère $D(A) = \sum_{k=1}^{K} D_k(A)$ où $D_k(A)$ est par exmple le nombre d'individus mal classés de la $k$ème feuille de l'arbre.

La construction de la séquence d'arbres emboités repose sur une pénalisation par la complexité de l'arbre :

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

Quand $\gamma = 0$, $A_{\mathsf{max}} = A_K$ minimise $C(A)$. En faisant croître γ, l'une des divisions de A_k, celle pour laquelle l'amélioration de D est la plus faible (inférieure à γ), apparaît comme superflue et les deux feuilles obtenues sont regroupées dans le noeud père qui devient terminal.

On peut représenter la décroissance ou ébouli de la déviance (ou du taux de mal classés) en fonction du nombre croissant de feuilles dans l'arbre.

6.4.2 Recherche de l'arbre optimal

On peut utiliser le graphe précédent en le lisant comme un ébouli de valeurs propres. Quand l'amélioration du critère est jugée trop petite on élague l'arbre.

6.5 Validation

Taux de bien/mal classés : validation croisée.
Lift chart : ratio des résultats obtenus avec et sans le modèle