Cours/TD Algorithme CART

Pour cet exercice, on va présenter l'algorithme CART (Classification and Regression Trees) dans le cadre de la classification binaire ($Y \in \{0,1\} := \mathcal{J}$) en utilisant un jeu de données fictif. On suppose que des données d'apprentissage ont été recueillies par un médecin sur un ensemble de 10 patients. Pour chacun des patients, le médecin a mesuré la pression artérielle du patient et indiqué des informations sur le patient (son âge, fumeur ou non et si il y avait des antécédents dans la famille du patient de maladie cardio-vasculaire). Suite à l'examen du patient, le médecin a indiqué si le patient présentait des risques de maladie cardio-vasculaire ou non. L'ensemble des données d'apprentissage est résumé dans le tableau ci-dessous.

patient	âge (X_1)	fumeur (X_2)	pression artérielle (X_3)	antécédents familiaux (X_4)	à risque (Y)
1	> 50	non	élevée	non	oui
2	≤ 50	oui	élevée	non	oui
3	> 50	oui	normale	oui	oui
4	> 50	non	élevée	oui	oui
5	> 50	oui	élevée	non	oui
6	≤ 50	non	élevée	non	non
7	≤ 50	oui	normale	oui	non
8	> 50	non	normale	oui	non
9	> 50	non	normale	oui	non
10	≤ 50	non	normale	non	non

On prendra dans la suite la convention que $Y = 1$ si la variable réponse est "oui" et 0 sinon. Le but de l'algorithme CART est de construire un arbre de classification comme celui de la Figure 1. Cet arbre est composé de nœuds et d'arrêtes. A partir de cet arbre de classification, on établit des prédictions en partant du haut de l'arbre et en descendant les arrêtes. On appelle feuille ou nœud terminal de l'arbre un nœud n'ayant pas de successeurs. En particulier, lorsqu'un nœud contient des observations provenant d'une seule classe, il est parfaitement homogène et on ne le découpe plus. Dans ce cas, on dit que le nœud est pur. L'arbre de la Figure 1 possède tous ses nœuds terminaux purs. La prédiction associée à un nœud terminal est alors effectuée selon la règle du vote à la majorité. C'est à dire que, par exemple, si le nœud contient une majorité d'observations labellisées 1 (dont le Y associé vaut 1) alors le nœud prédit le label 1. Pour prédire si un patient est à risque ou non à partir de cet arbre de classification, il suffit donc de partir du haut de l'arbre (la racine de l'arbre) et de descendre les arrêtes jusqu'à arriver à une feuille et de lire sa prédiction.
On note $D_n = \{(X_i, Y_i), \quad i = 1, \ldots, n\}$ notre jeu de données dont on dispose pour la construction de l'arbre (dans le cadre de l'exercice, on a $n = 10$) de la Figure 1.
La construction d'un arbre de classification CART s'effectue en trois étapes :

- Construction d'un arbre complet T^c
- Élagage de l'arbre T^c (construction d'une suite de sous arbres de T^c)
- Sélection par validation croisée de l'arbre final T^* parmis la suite de sous arbres construits lors de l'étape d'élagage.

Construction de l'arbre complet T^c : Pour la première étape, c'est le critère d'homogénéité qui guide la construction de l'arbre complet. C'est à dire que l'on va chercher les découpes permettant d'obtenir les nœuds les plus homogènes possibles. On commence par le noeud racine, que l'on va noter ici t, et on va considérer des découpes de la forme $X_i > \alpha$, ou α est un réel lorsque la variable X_i est quantitative (variable numérique) ; et $X_i = 1$ si la variable est binaire ($X_i \in \{0,1\}$ comme dans l'exemple). Dans la suite, on note $\mathcal{D}$ l'ensemble des découpes possibles pour le noeud t et on considère $D \in \mathcal{D}$ une découpe. On note t_D^g et t_D^d les deux noeuds issus de t et induit par la découpe D. Par exemple si D est de la forme $X_i = 1$, t_D^g est constitué des observations du noeuds t satisfaisant $X_i = 1$ et t_D^d des observations de t satisfaisant $X_i = 0$. On souhaite déterminer la découpe

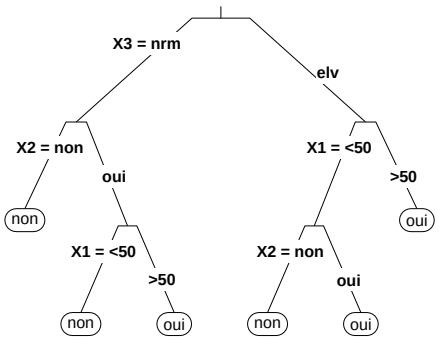

FIGURE 1 – Arbre de classification maximal T^c. Les abréviations nrm et elv sont utilisées pour normale et élevée.

D pour laquelle le gain d'homogénéité est maximal. L'homogénéité est mesuré par l'indice de Gini, défini comme suit : dans le cadre de la classification binaire, l'indice de Gini associé à un noeud t est défini par

$$I(t) = 1 - \sum_{j \in \mathcal{J}} \hat{p}_j^2(t),$$

où $\hat{p}_j \in [0,1]$ est la proportion d'observations dans le noeud t associées à la classe j. Pour mesurer le gain d'homogénéité, on considère alors la quantité suivante

$$\Delta(t, D) = I(t) - \left(p_g I(t_D^g) + p_d I(t_D^d)\right),$$

où p_g (p_d respectivement) est la proportion d'observations du noeud t appartenant au noeud t_D^g (t_D^d respectivement). Ainsi l'importance dans le calcul du gain d'homogénéité de chacun des deux noeuds t_D^g et t_D^d est pondérée par le nombre d'observations qu'ils contiennent. Enfin la découpe optimale D^* est naturellement définie comme le maximiseur de $\Delta(t, D)$

$$D^* = \arg\max_{D \in \mathcal{D}} \Delta(t, D).$$

Ensuite, on itère le procédé décrit ci-dessus pour chacun des noeuds créés jusqu'à un critère d'arrêt choisi au préalable. Par exemple, on peut considérer le critère consistant à arrêter la découpe d'un nœud si il contient moins qu'un nombre d'observations fixé à l'avance. Par ailleurs, on ne découpe pas un nœud pur. Enfin, pour conclure, on pourrait penser à considérer le risque de classification plutôt que l'indice de Gini pour déterminer les découpe optimale. La raison pour laquel on ne l'utlise pas est que le risque de classification donne des nœuds moins homogènes que l'indice de Gini. Or un nœud homogène est un nœud qui sera très peu découpé dans la suite de l'algorithme (gain en terme de complexité pour l'algorithme).

1. On considère donc l'arbre de classification maximal T^c de la Figure 1. Soit t un nœud de l'arbre.
   (a) Combien l'arbre donné en Figure 1 a-t-il de feuilles ?
   (b) Donner le graphe de la fonction $x \mapsto x(1-x)$.
   (c) Montrer que l'indice de Gini $I(t)$ du nœud t peut s'écrire :
   $$I(t) = 2\hat{p}_1(t)(1 - \hat{p}_1(t)).$$
   (d) Que peut-on dire lorsque $I(t) = 0$ (le nœud t est alors dit pur) ? $I(t) = 1/2$? En particulier, que vaut l'erreur de classification $R(t)$ dans chacun de ces cas ?
2. Justifier par le calcul que la première variable utilisée pour la construction de l'arbre est la variable X_3.
3. On considère la donnée suivante associée à un nouveau patient :
   âge : ≤ 50 *; fumeur : oui ; pression artérielle : élevée ; antécédents familiaux : oui.*
   À l'aide de l'arbre T^c déterminer si le patient est à risque ou non.

Élagage de l'arbre complet L'arbre complet T^c construit dans la première étape n'est pas forcément un bon classifieur. En effet bien que l'erreur de classification commise par T^c sur l'échantillon d'apprentissage est petite (et même nulle si ses nœuds terminaux sont tous purs), il peut avoir une mauvaise erreur de généralisation, c'est à dire qu'il classe mal de nouvelles observations. Ceci est du à sa trop grande complexité, on dit qu'il "colle" trop au données. Dans ce cas on parle de phénomène de sur apprentissage. À contrario, si on considère la racine de l'arbre T^c comme classifieur, on parle de sous apprentissage (on a rien appris des données). L'objectif de cette étape et de la suivante est donc de déterminer un sous-arbre de T^c réalisant un compromis entre performance et complexité (on peut rapprocher cela du compromis biais variance). Pour mieux comprendre le phénomène de sur apprentissage, on peut se référer à l'exemple de l'algorithme des 1 plus proches voisins où l'erreur de classification est nul sur les données d'apprentissage mais l'erreur de généralisation sur données tests est mauvaise (en particulier, ce n'est pas un algorithme consistant). Dans la suite, on note $\mathcal{T}$ l'ensemble des sous-arbres de T^c.

Pour réaliser le compromis entre performance et complexité, un principe général en statistique est de "pénaliser" un prédicteur par sa complexité. Pour un arbre T, sa complexité est mesuré par son nombre de nœuds terminaux (noté $|\tilde{T}|$). Plus celui-ci est grand plus l'arbre est complexe. Son risque de classification $R(T)$ est mesuré comme le nombre d'erreurs commises sur l'échantillon d'apprentissage par chacun des nœuds terminaux divisé par le nombre total d'observations. Comme vue en question 1) pour un nœud terminal t, on détermine la classe majoritaire associée à t, et le nombre d'erreurs commises sur ce nœud est donc le nombre d'observations issues de la classe minoritaire. Pour un paramètre $\alpha > 0$ dit de pénalisation (ou température), on considère alors le critère pénalisé suivant

$$\mathrm{Crit}_\alpha(T) = R(T) + \alpha|\tilde{T}|, \;\; T \in \mathcal{T}$$

et l'on considère alors T_α définit comme le minimiseur de Crit_α sur l'ensemble des sous-arbres de T^c.

$$T_\alpha = \arg\min_{T \in \mathcal{T}} \mathrm{Crit}_\alpha(T).$$

On constate que lorsque α est grand, l'arbre sélectionné T_α aura une complexité faible alors que si α, T_α se rapprochera de l'arbre complet en terme de complexité. Évidemment, le critère Crit_α réalisant le compromis entre performance et complexité dépend du paramètre α qu'il faut choisir.

Pour déterminer l'arbre final, on ne peut pas explorer tout les sous-arbres de T^c. Ce serait trop long d'un point de vue algorithmique. L'objectif de cette étape est donc de déterminer une suite de sous-arbres $T_0, \ldots, T_K$ parmi lesquels sera sélectionné l'arbre final. Pour cela, on détermine une suite $\alpha_0, \ldots, \alpha_K$ de paramètres de pénalisation (on ne détaille pas le choix des paramètres α_i). Finalement, on construit la suite $(T_0, \ldots, T_K)$, en posant

$$T_i = T_{\alpha_i}.$$

4. Donner une formule simple pour calculer le risque de classification $R(T)$ d'un arbre. Que vaut $R(T^c)$ sur l'arbre de la Figure 1 ? En déduire $\mathrm{Crit}_\alpha(T^c)$ en fonction de α.
5. On considère maintenant les deux-sous arbres de la Figure 2 ainsi que le sous-arbre uniquement constitué de la variable X_3. Donner les prédictions associées à chaque noeud terminal de ces trois arbres. On remarquera que dans le cas où il n'y a pas de vote majoritaire, l'attribution de la prédiction "oui" ou "non" donne le même taux de mauvaise classification pour la feuille correspondante. Calculer $\mathrm{Crit}_\alpha(T)$ en fonction de α pour ces trois sous arbres.
6. Montrer que le sous-arbre de droite de la Figure 2 ne peut pas faire partie de la suite des sous-arbres parmi lesquels sera sélectionné l'arbre final.
7. Pour $\alpha = 0$ quel est l'arbre T_0 sélectionné ? A partir de quelles valeurs de α l'arbre sélectionné T_α ne peut-il pas être l'arbre maximal ?

Sélection de l'arbre final T^* L'arbre final T^* est alors choisi par validation croisée dans l'ensemble $\{T_0, \ldots, T_K\}$. Il faut noter qu'un arbre de décision a pour principal avantage d'offrir une règle de décision très facilement interprétable et par ailleurs, on peut la représenter de manière visuelle. Il suffit alors de lire l'arbre pour comprendre l'apport des covariables à la règle de décision. De plus, la construction de l'arbre final T^* ne repose sur aucune hypothèse de modélisation, au contraire par exemple de la régression logistique (on parle d'estimateur non paramétrique). Néanmoins, les arbres de décision souffre d'un défaut majeur qui est leur manque de stabilité, une légère perturbation de l'échantillon d'apprentissage peut conduire à un arbre très différent. Par ailleurs, il ne font pas parti des règles de prédictions les plus performantes mais sont utilisés comme ingrédients de base dans des algorithmes beaucoup plus performants (mais moins interprétable) comme le *boosting* ou les *forêts aléatoires*.

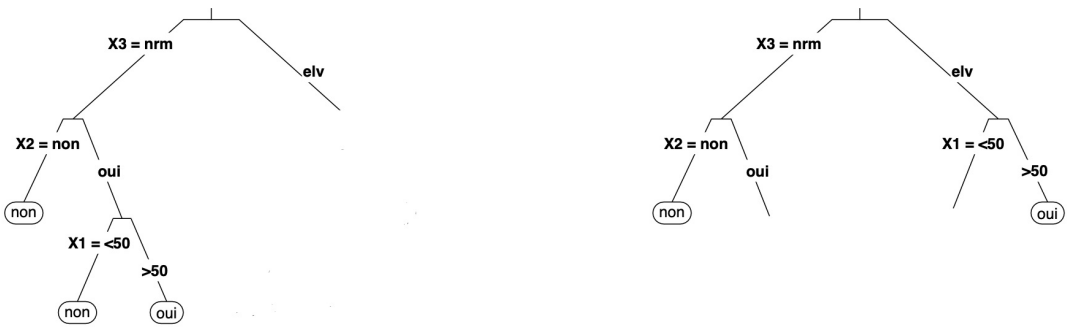

FIGURE 2 – Deux sous-arbres obtenus à partir de l'arbre maximal de la Figure 1.

8. Dans cette question les données sont modélisées comme suit :
   $\mathbf{X} = (X_1, X_2, X_3, X_4) \in \{0,1\}^4$ et $Y \in \{0,1\}$ avec la convention
   $> 50 =$ elevee $=$ oui $= 1$ et $\leq 50 =$ normale $=$ non $= 0$.
   On définit $\forall \mathbf{x} = (x_1, x_2, x_3, x_4) \in \{0,1\}^4$, $\eta(\mathbf{x}) = P(Y = 1 | \mathbf{X} = \mathbf{x})$.
   On suppose de plus que $\forall \mathbf{x} \in \{0,1\}^4$ et $y \in \{0,1\}$:

$$P(\mathbf{X} = \mathbf{x} | Y = y) = \prod_{i=1}^{4} P(X_i = x_i | Y = y)$$

   (a) Montrer que pour tout $\mathbf{x} \in \{0,1\}$, on a

$$\eta(\mathbf{x}) \geq 1/2 \text{ ssi } P(\mathbf{X} = \mathbf{x} | Y = 1) P(Y = 1) \geq P(\mathbf{X} = \mathbf{x} | Y = 0) P(Y = 0)$$

   (b) Pour $y \in \{0,1\}$ et $i = 1, 2, 3, 4$, Proposer, à l'aide du Tableau , une estimation de $P(Y = y)$, $P(X_i = 1 | Y = y)$.

   (c) À l'aide des deux questions précédentes et du tableau, déterminer une estimation de $s^*(\mathbf{x})$ où s^* est le classifieur de Bayes et $\mathbf{x} = (0, 1, 1, 1)$.