Exercice de mise en œuvre des arbres de décision

Fabrice Rossi

On étudie dans cette section un ensemble de 366 patients atteints de six maladies de peau différentes. Chaque patient est décrit par 35 variables, l'objectif étant de classer automatiquement un patient dans la classe de sa maladie de peau. Il y a donc 6 classes, la classe d'un patient étant donnée par la valeur de la 35ème variable (un entier de 1 à 6). Les 34 autres variables peuvent prendre 4 valeurs distinctes, les entiers 0, 1, 2 et 3.

Question 1 *On construit un arbre de décision sur l'ensemble des patients. L'arbre complet obtenu comprend 21 feuilles. On simplifie arbitrairement l'arbre à 4 feuilles, ce qui donne l'arbre de la figure 1.*

La table 1 donne la description d'un patient par classe en la réduisant aux trois variables utilisées dans l'arbre. Donner pour chaque patient la classe prédite par l'arbre.

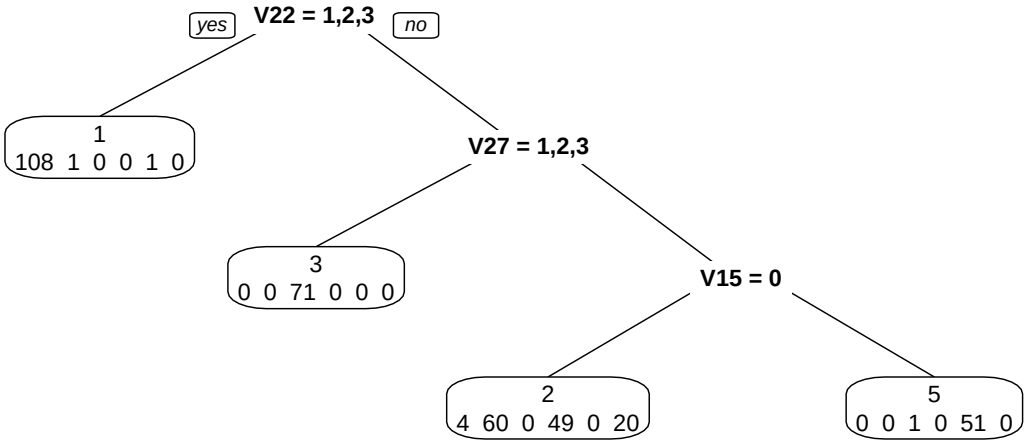

FIGURE 1 – Arbre de classification à 4 feuilles. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». Le chiffre indiqué sur la première ligne de chaque feuille donne la classe associée à la feuille. Les six nombres de la deuxième ligne d'une feuille indiquent les effectifs des 6 classes dans la feuille, dans l'ordre.

Question 2 *Déduire de la figure 1 la matrice de confusion de l'arbre simplifié qu'elle représente. On donnera le résultat sous la forme utilisée par la table 2.*

Question 3 *L'arbre complet avec 21 feuilles a la matrice de confusion sur l'ensemble des patients donnée par la table 2. Commentez les résultats obtenus.*

Question 4 *La figure 2 représente le nombre d'erreurs de classement commises par l'arbre en fonction du nombre de feuilles conservées, sur l'ensemble d'apprentissage (ligne en tirets) et grâce à une procédure de validation croisée à 10 blocs (ligne pleine). Combien faut-il conserver de feuilles d'après cette figure pour obtenir le meilleur modèle ? En quel sens le modèle retenu est-il le meilleur ?*

	V15	V22	V27	Classe
2	0	2	0	1
21	0	0	0	2
3	0	0	2	3
9	0	0	0	4
7	3	0	0	5
211	0	0	0	6

TABLE 1 – Description d'un patient de chaque classe pour les variables utilisées dans l'arbre de la figure 1

	1	2	3	4	5	6
1	112	0	0	0	0	0
2	0	61	0	0	0	0
3	0	0	72	0	0	0
4	0	0	0	49	0	0
5	0	0	0	0	52	0
6	0	0	0	0	0	20

TABLE 2 – Matrice de confusion de l'arbre complet : chaque ligne correspond à la classe réelle du patient, chaque colonne à la classe prédite par l'arbre.

Question 5 *Après simplification optimale de l'arbre, on obtient la matrice de confusion de la table 3 sur l'ensemble d'apprentissage. Commentez les résultats obtenus.*

Question 6 *On découpe aléatoirement l'ensemble des données en deux sous-ensembles disjoints de même taille A et B, en respectant les proportions des six classes. On construit un arbre complet sur l'ensemble A, puis on le simplifie comme à la question 4. La table 4 donne les matrices de confusion de cet arbre sur les ensembles A et B.*

Expliquez pourquoi on respecte les proportions des classes dans la partition. Commentez les matrices de confusion et la procédure qui permet de les obtenir, en comparant notamment aux résultats obtenus à la question 5.

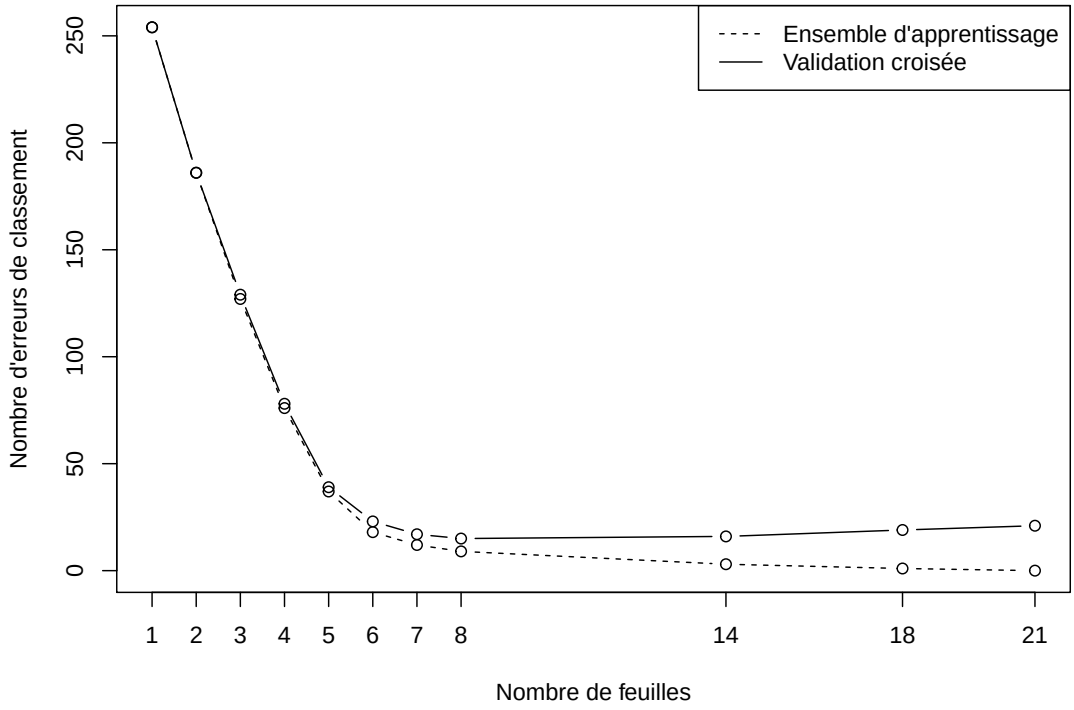

FIGURE 2 – Évolution du nombre d'erreurs de classement en fonction du nombre de feuilles dans l'arbre, sur l'ensemble d'apprentissage et estimé par validation croisée.

	1	2	3	4	5	6
1	111	1	0	0	0	0
2	1	58	0	1	0	1
3	0	0	71	0	1	0
4	0	3	0	46	0	0
5	1	0	0	0	51	0
6	0	0	0	0	0	20

TABLE 3 – Matrice de confusion de l'arbre simplifié : chaque ligne correspond à la classe réelle du patient, chaque colonne à la classe prédite par l'arbre.

Ensemble A

	1	2	3	4	5	6
1	56	0	0	0	0	0
2	0	31	0	0	0	0
3	0	0	36	0	0	0
4	0	1	0	24	0	0
5	0	0	0	0	26	0
6	0	0	0	0	0	10

Ensemble B

	1	2	3	4	5	6
1	53	2	0	1	0	0
2	0	27	0	2	0	1
3	0	1	33	2	0	0
4	0	5	0	19	0	0
5	0	0	0	0	26	0
6	1	0	0	0	0	9

TABLE 4 – Matrices de confusion de l'arbre obtenu à la question 6 : chaque ligne correspond à la classe réelle du patient, chaque colonne à la classe prédite par l'arbre.