Université Paris-Saclay-Faculté de médecine
Master 1 de santé publique : introduction aux sciences de données

Session du 05 juin 2024 de 15h à 16h30.

I. Compréhension de cours

A. Laquelle des affirmations suivantes est vraie concernant le modèle des k plus proches voisins ?

(a) Il est toujours préférable de choisir un k plus petit.
(b) Un modèle à k voisins les plus proches a exactement k paramètres entraînables.
(c) La précision d'apprentissage d'un modèle des 3 plus proches voisins est généralement supérieure à celle d'un modèle de 1 plus proche voisin.
(d) La frontière de décision d'un modèle de k plus proches voisins est linéaire.
(e) Plus k augmente, plus le biais (au sens de la décomposition biais-variance) augmente.

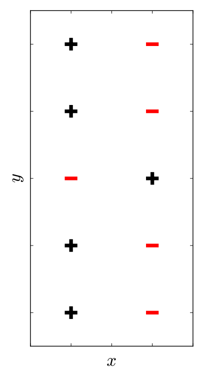

FIGURE 1 – Jeu de données avec deux variables explicatives x et y. La variable à expliquer est le signe $-$ et $+$.

B. Reprendre la figure 1, tracer la frontière de classement (on peut utiliser des hachures pour distinguer une classe d'une autre) des deux modèles k plus proches voisins avec $k = 1$ et $k = 3$ (**plus difficile**).

II. Arbres de décision

A. Tracer l'arbre de décision correspondant à la partition illustrée sur la partie gauche de la figure 2. Le chiffre à l'intérieur de chaque case indique la moyenne de Y à l'intérieur de la région correspondante.

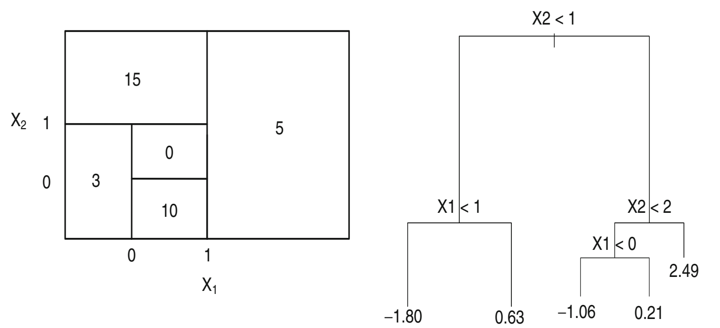

FIGURE 2 – Attention : les deux figures représentent deux modèles différents. À gauche : une partition de l'espace des variables explicatives. À droite : un arbre de décision.

A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F

FIGURE 3 – Pseudo jeu de données composé de trois variables explicatives binaires (A, B et C) et une variable binaire à expliquer Y.

B. Tracer une partition similaire à la partition à droite de la figure 2 à partir de de l'arbre illustré à droite de la figure 2. Les régions disjointes de la partition représentent les feuilles de l'arbre. Indiquer la moyenne de Y à l'intérieur de chaque région.

C. Proposer un arbre qui classe parfaitement le pseudo jeu de données du tableau de la figure 3 (aucun calcul n'est nécessaire pour justifier votre réponse).

D. Considérons les trois variables A, B et C du pseudo jeu de données du tableau de la figure 3 comme des variables binaires où la modalité F correspond 0 et la modalité T correspond à 1. Supposons que nous souhaitons apprendre une fonction qui compte le nombre de 1 dans un vecteur de variables binaires.

- **a.** Proposer un arbre de décision qui permet de compter (sans erreur) le nombre de 1 dans un vecteur de variables explicatives binaires de dimension 3.
- **b.** De combien de feuilles a-t-on besoin ?

c. De combien de feuilles a-t-on besoin pour compter (sans erreur) le nombre de 1 dans un vecteur de variables explicatives binaires de dimension d ?

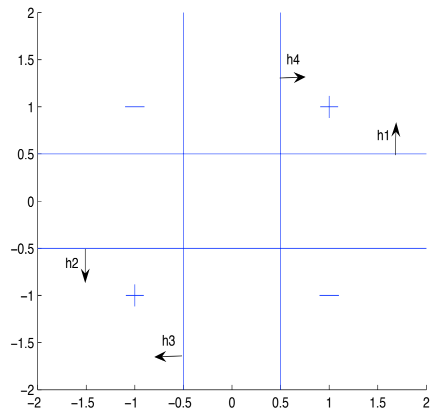

FIGURE 4 – Un exemple de jeu de données avec deux variables x_1 et x_2. Les règles faibles $h_1, \ldots, h_4$ sont des arbres à deux feuilles .

II. Un peu de boosting

A. Une règle de classification combinée $H_M(x)$ (obtenue par AdaBoost par exemple) est donnée par

$$H_M(x) = \text{sign}\Big(\sum_{m=1}^{M} \alpha_m h_m(x)\Big),$$

où h_m est une règle de classification faible ajustée à l'itération m. Supposons que nous souhaitons utiliser les 4 règles faibles indiquées (la flèche de chaque règle faible indique la région correspondante à la classe + ($i.e$ $h_m(x) > 0$)) sur la figure 4 pour construire une règle de classification combinée $H_4(x)$. Montrer qu'il n'y a pas de poids $\alpha_1, \alpha_2, \alpha_3$ et α_4 qui permettent à $H_4(x)$ de classer parfaitement les 4 données de la figure 4.