Université Paris-Saclay : faculté de médecine
Master : santé publique
Introduction aux data sciences

Session du 08 Juin 2021 à 14h.

I. Arbres : représentations graphiques

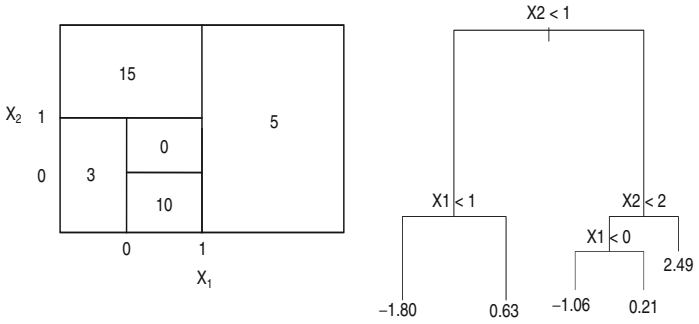

FIGURE 1 – Attention : les deux figures représentent deux modèles différents. À gauche : une partition de l'espace des variables explicatives. À droite : un arbre de décision.

A. Tracer l'arbre de décision correspondant à la partition illustrée sur la partie gauche de la figure 1. Le chiffre à l'intérieur de chaque case indique la moyenne de Y à l'intérieur de la région correspondante.

B. Créer un diagramme (une partition) similaire à la partition à droite de la figure 1 à partir de de l'arbre illustré à droite de la figure 1. Les régions disjointes de la partition doivent représentées les feuilles de l'arbre et contenir la moyenne de Y à l'intérieur de chaque région.

II. Arbres : généralités

A. Supposons que nous disposons d'échantillons indépendants contenants deux classes rouge et verte (modalités de la variable réponse Y). Nous ajustons un arbre de classification sur chaque échantillon et pour une valeur test de $X = x$, nous produisons 10 approximations de la probabilité $\mathbb{P}\big(\text{la classe est Rouge} \mid X = x\big)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, \text{ et } 0.75.$$

Il existe deux méthodes couramment utilisées pour combiner ces 10 résultats en une seule prédiction de classe. La première méthode consiste à prendre comme prédiction finale, le vote majoritaire simple. La deuxième approche consiste à classer en fonction de la probabilité moyenne des 10 approximations. Sur cet exemple, quelle est la prédiction finale selon chacune de ces deux approches ?

B. Rappeler le principe de fonctionnement de l'algorithme CART pour la régression. Décrire et détailler les étapes de création de la racine de l'arbre et décrire la procédure de prédiction à partir d'un arbre de décision.

C. Donner un avantage et un inconvénient de la modélisation par un arbre de décision unique.

III. Régression linéaire : complexité d'un modèle

A. Supposons que nous disposons d'un de jeu de données de $n = 100$ observations contenant une seule variable explicative et une variable réponse quantitative. Nous ajustons un modèle de régression linéaire ainsi qu'un modèle de régression cubique, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

(a) Supposons que la vraie relation entre X et Y est linéaire, $Y = \beta_0 + \beta_1 X + \varepsilon$. Considérons la somme des carrés résiduels (RSS) calculée sur le jeu de données d'apprentissage associée à la régression linéaire ainsi que RSS calculée sur le jeu de données d'apprentissage associée à la régression cubique. Les deux quantités précédentes sont-elles égales ? l'une est supérieure à l'autre ? ou aucune des deux affirmations précédentes n'est vraie ? Justifiez votre réponse.

(b) Répondre à la question précédente en considérant des quantités RSS calculées sur un jeu de données test.

(c) Supposons que la vraie relation entre X et Y est non linéaire mais nous n'avons aucune information sur l'écart de cette relation à la linéarité. Considérons la somme des carrés résiduels (RSS) calculée sur le jeu de données d'apprentissage associée à la régression linéaire ainsi que RSS calculée sur le jeu de données d'apprentissage associée à la régression cubique. Les deux quantités précédentes sont-elles égales ? l'une est supérieure à l'autre ? ou aucune des deux affirmations précédentes n'est vraie ? Justifiez votre réponse.

(d) Répondre à la question précédente en considérant des quantités RSS calculées sur un jeu de données test.

IV. Bagging et forêts aléatoires

A. Rappeler en quelques mots le principe d'agrégation d'arbres par *bagging* en régression et classification.

(a) Une forêt aléatoire est-elle un cas particulier d'agrégation d'arbres par *bagging* ou l'inverse ? Donner un cas particulier où les deux modèles coïncident.

(b) Donner deux exemples de modèles à faible biais et grande variance.