SY19 – A18
TP 1: Prise en main de R, k plus proches voisins, décomposition biais-variance

1 Prise en main de R

Charger le livre "An Introduction to Statistical Learning" depuis la page Moodle. Si vous n'avez aucune connaissance préalable de R, lisez l'introduction de la section 3.3 (page 42 et suivantes), en testant en même temps les commandes R présentées.

2 Régression par la méthode des k plus proches voisins

On cherche dans cet exercice à prédire le niveau de PSA (prostate-specific antigen) à partir de variables cliniques.

1. Chargez les données `prostate.data`.
2. Cherchez graphiquement quelles variables semblent expliquer la variable `lpsa` (logarithme de la concentration en PSA).
3. Mettre en œuvre sur ces données la régression par la méthode des k plus proches voisins, à l'aide de la fonction `knn.reg` du package `FNN`. On utilisera comme prédicteurs les variables `lcavol`, `lweight`, `age` et `lbph`. L'objectif est de prédire la variable `lpsa` pour les données de test, en utilisant les données d'apprentissage. (On utilisera la partition indiquées dans les données par la variable logique `train`).
4. Représenter graphiquement l'erreur quadratique moyenne estimée sur l'ensemble d'apprentissage à partir de k. Pour quelle valeur de k obtient-on l'erreur la plus faible ?

3 Etude du compromis biais-variance

On considère le modèle suivant :

$$Y = 1 + 5X^2 + \varepsilon \tag{1}$$

avec $X \sim \mathcal{U}([0,1])$ et $\epsilon \sim \mathcal{N}(0, \sigma^2)$, où $\sigma = 0.5$. Soit $\widehat{f}_k(x)$ l'estimation de la fonction de régression $f(x) = 1 + 5x^2$ obtenue en calculant la moyenne des y_i pour les k plus proches voisins de x. On rappelle la formule de la décomposition biais-variance :

$$\mathbb{E}\left[(\widehat{f}_k(x) - Y)^2 \mid X = x\right] = \mathrm{Var}\left(\widehat{f}_k(x)\right) + \left(\mathbb{E}[\widehat{f}_k(x)] - f(x)\right)^2 + \mathrm{Var}(\varepsilon \mid X = x). \quad (2)$$

1. Expliquez les différents termes de cette formule et démontrez la.
2. Vérifiez la formule (2) par simulation, en générant aléatoirement des ensembles d'apprentissage de taille $n = 50$. Pour une certaine valeur de x, on tracera les différents termes de la formule en fonction de k, pour k compris entre 1 et 40. Qu'observe-t-on ? Répétez l'expérience pour différentes valeurs de n et commentez les résultats.