TP : Généralités

Vous pouvez faire les exercices dans le langage de votre choix.

1 Exercice 1 : Nettoyage et exploration de données

1. Télécharger le jeu de données suivant : lien
2. Nettoyer le jeu de données. En particulier, on s'intéressera aux points suivant :

— encodage des valeurs manquantes ;
— gestion des valeurs extrêmes ;
— gestion des valeurs abberantes ;
— gestion des doublons ;
— harmonisation des dates ;
— suppression des tirets, points, … dans les numéros de téléphone.
— gestion des titres (Mr., Ms., Dr., …) dans les noms ;
— harmonisation des pays ;
— conversion des Oui/Non en TRUE/FALSE ;
— etc.

3. Faire une exploration unidimensionelle/bidimensionelle. En particulier, on pourra faire :

— graphiques des données quantitatives ;
— calcul de corrélations ;
— tableaux de fréquences pour les variables qualitatives ;
— etc.

2 Exercice 2 : Compromis biais-variance

Dans cet exercice, on se propose d'illustrer le compromis biais-variance à l'aide de données simulées et d'un modèle de régression. On fait l'hypothèse que le vrai modèle de nos données est

$$f(x) = 3 + 8x + 2x^2.$$

1. Simuler un ensemble de données (X, Y) tel que $Y = f(X) + \epsilon$, avec $X \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ et $\sigma^2 = 5$.
2. Ajuster un modèle linéaire de la forme $Y = \beta_0 + \beta_1 X$. En R, la fonction standard est `lm`. En Python, vous pouvez utiliser la fonction `ols` du package `statsmodels`.
3. Calculer $\hat{Y}$ en utilisant X généré à la question 1.
4. Calculer l'erreur quadratique moyenne sur le jeu d'entraînement.
5. Simuler un jeu de données de validation et calculer l'erreur quadratique moyenne, le biais et la variance du modèle sur ce jeu de validation.
6. Refaire les questions 2 à 5, mais avec en ajustant un modèle linéaire avec un polynôme d'ordre 2.
7. Refaire les questions 2 à 5, mais avec en ajustant un modèle linéaire avec un polynôme d'ordre 10.
8. Conclure quant au compromis biais-variance et la complexité du modèle.

3 Exercice 3 : Validation croisée

En pratique, nous ne connaissons pas la vraie fonction $f(x)$. Pour évaluer la qualité de notre modèle, on peut utiliser la validation croisée. Notez que le but de cet exercice est de faire votre propre code et non d'utiliser des fonctions déjà faites.

1. Simuler un jeu de données en utilisant le modèle de l'exercice 2.
2. Faire une validation croisée avec $K = 3$ pour évaluer le modèle de régression linéaire simple (sans polynôme d'ordre supérieur) en utilisant l'erreur quadratique moyenne.
3. Faire la même chose avec le modèle de régression linéaire avec un polynôme d'ordre 2.
4. Faire la même chose avec le modèle de régression linéaire avec un polynôme d'ordre 10.
5. Conclure quant à l'utilisation de la validation croisée.