DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL

MTH6312 - MÉTHODES STATISTIQUES D'APPRENTISSAGE

Devoir nº 3 - Automne 2024

Date de remise : 14 novembre avant 23h55 (en pdf dans Moodle)

DIRECTIVES :

- ✓ Inclure dans votre rapport le code R (ou Python) que vous avez utilisé. Dans votre code le germe utilisé dans les questions de ré-échantillonnage est votre matricule, c-à-d `set.seed(matricule)`.
- ✓ Lors de la correction, il sera tenu compte de la clarté des démarches ainsi que de la qualité de la présentation du rapport.

QUESTION Nº 1 (10 points). Pour cette question, comme au devoir 2, vous devez d'abord obtenir vos données personnalisées (*mondata1*) **en fonction de votre matricule** (voir les instructions à cet effet sur le site). Il s'agit d'un échantillon aléatoire de 500 observations de la base de données *Wage* du *package* ISLR2.

Vous disposez ainsi de 500 observations sous la forme $\{(\mathbf{x}_i, y_i), i = 1, \ldots, 500\}$, où y_i représente l'état de santé du travailleur (`health`) codée 1 (très bon), 0 (bon ou moins); $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$, où x_{i1} est l'âge (`age`) du travailleur et x_{i2} son salaire (`wage`).

Pour les données décrites ci-dessus, après avoir déterminé le degré de flexibilité approprié de chaque méthode, on veut sélectionner la meilleure méthode de classification parmi les suivantes : le KNN, la régression logistique, l'analyse discriminante linéaire et l'analyse discriminante quadratique.

a) KNN.

1. En utilisant les techniques de validation croisée *«LOOCV»* et *«5-Fold CV»* sur les 500 observations, estimer le taux d'*erreur test* pour différentes valeurs du nombre de voisins, K, avec $K = 1, 2, \ldots, 180$. Tracer la courbe du taux d'erreur en fonction de $1/K$.
2. Compte tenu des résultats ci-dessus, quelle valeur du nombre de voisins K devrait-on utiliser pour la classification des données du contexte par le KNN? Justifier brièvement.

b) Régression logistique.

Dans les équations suivantes $p(\mathbf{x})$ représente $p(\mathbf{x}; \boldsymbol{\beta}) = P(Y = 1 \mid X = \mathbf{x})$, la probabilité que le travailleur soit en très bonne santé étant donné les mesures $\mathbf{x} = (x_1, x_2)^\top$ observées.

On envisage deux modèles de régression logistique dont les équations sont :

$$\text{Modèle 1 :} \quad \ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{Modèle 2 :} \quad \ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2.$$

1. En utilisant les techniques de validation croisée *«LOOCV»* et *«5-Fold CV»* sur les 500 observations, estimer le taux d'*erreur test* pour chacun des 2 modèles.
2. Compte tenu des résultats ci-dessus, lequel des 2 modèles de régression logistique devrait-on utiliser pour la classification des données du contexte ? Justifier brièvement.

c) Analyse discriminante.

1. En utilisant les techniques de validation croisée *«LOOCV»* et *«5-Fold CV»* sur les 500 observations, estimer le taux d'*erreur test* de l'analyse discriminante linéaire (LDA) et celui de l'analyse discriminante quadratique (QDA).
2. Compte tenu des résultats ci-dessus, laquelle des deux analyses discriminantes devrait-on utiliser pour la classification des données du contexte ? Justifier brièvement.

d) Résumé graphique et comparaison des méthodes.

Tracer le nuage des 500 points (2 couleurs de votre choix) et ajouter au graphique les courbes (trois en tout, similaires à celles de la figure 5.7 page 207 dans ISLr) séparant les deux classes dans chacun des cas suivants :

- le KNN (avec la valeur optimale retenue du nombre de voisins K).
- le modèle de régression logistique retenu (parmi les deux modèles considérés) ;
- l'analyse discriminante retenue (LDA ou QDA).

Comparer les trois méthodes en utilisant leurs matrices de confusion des indices de performance.

QUESTION N° 2 (5 points). Pour cette question on utilise les données `iris` (disponibles dans R). Ces données sont de la forme $\{\mathbf{x}_i, i = 1, \ldots, n\}$ où $n = 150$ et chaque x_i est de dimension $p = 4$. On considère que chaque $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top, i = 1, \ldots, n$ est une observation d'un vecteur aléatoire $X = (X_1, X_2, X_3, X_4)^\top$, où les variables X_1 et X_2 représentent la longueur et la largeur des sépales, X_3 et X_4 la longueur et la largeur des pétales de fleurs. La distribution des probabilité exacte du vecteur X est inconnue. On s'intéresse ici à l'estimation du paramètre θ défini par

$$\theta = \mathbb{E}\left(\min\left\{X_2 + \log(X_1),\ X_1 + X_3 - 2X_4,\ \exp\left\{-|X_1 - X_4|\right\},\ X_2 + 3X_3\right\}\right).$$

a) (1 point) Donner l'expression d'un estimateur $\hat{\theta}$ de θ en fonction des n observations.

a) (2 points) En utilisant la technique de ré-échantillonnage *«Bootstrap»* (fonction `boot()`) avec 3500 répétitions, donner une estimation ponctuelle $\hat{\theta}$. Donner ensuite une estimation du biais et une estimation de l'écart type (erreur-type) de $\hat{\theta}$.

b) (2 points) Déduire des résultats qui précèdent un intervalle de confiance pour θ au niveau de confiance 95%. Commenter brièvement.

QUESTION N° 3 (5 points). Exercice n° 8 page 285 ISLr (An Introduction to Statistical Learning).
Générer les données (*mondata3*) **en fonction de votre matricule**, voir les instructions à sur le site :

1. générer les observations x_i selon une normale $N(\mu, \sigma^2)$ avec μ et σ de votre choix ;
2. générer les erreurs ε_i selon une normale $N(0, \sigma_\varepsilon^2)$ avec un σ_ε de votre choix.