DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL

MTH6312 - MÉTHODES STATISTIQUES D'APPRENTISSAGE

Devoir nº 3 - Automne 2018

Date de remise : 9 novembre avant 23h55 (en pdf dans Moodle)

DIRECTIVES :

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

QUESTION Nº 1 (10 points). Considérons de nouveau le contexte et les données de la question nº 2 du devoir nº 2 (voir fichier *Equipement.csv* sur le site du cours). Les données sont de la forme $\{(\mathbf{x}_i, y_i), i = 1, \ldots, 250\}$, où $\mathbf{x}_i = (x_{i1}, x_{i2})^\top$, x_{i1} étant la mesure de X_1, x_{i2} celle de X_2, et y_i représente l'état réel de l'équipement lors de la i^{e} observation. Deux modalités sont utilisées pour y_i : (D) présence d'anomalies, ou (N) absence d'anomalies.

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 : le KNN, la régression logistique, l'analyse discriminante linéaire et l'analyse discriminante quadratique.

a) KNN.

1. En utilisant la technique de validation croisée *«LOOCV»* sur les 250 observations, estimer le taux d'*erreur test* pour différentes valeurs (une dizaine) du nombre de voisins, k.
2. Reprendre la question ci-dessus en utilisant une des techniques de validation croisée *«5-Fold CV»* ou *«10-Fold CV»* de votre choix. Justifier brièvement votre choix.
3. 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 $p(\mathbf{x}; \boldsymbol{\beta}) = P(Y = D \mid X = \mathbf{x})$, la probabilité que l'individu soit en défaut de paiement étant donné son revenu x_1 et le montant dû x_2.

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 la technique de validation croisée *«LOOCV»* sur les 250 observations, estimer le taux d'*erreur test* pour chacun des 2 modèles.
2. Reprendre la question ci-dessus en utilisant une des techniques de validation croisée *«5-Fold CV»* ou *«10-Fold CV»* de votre choix. Justifier brièvement votre choix.
3. 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 la technique de validation croisée *«LOOCV»* sur les 250 observations, estimer le taux d'*erreur test* de l'analyse discriminante linéaire (LDA) et celui de l'analyse discriminante quadratique (QDA).
2. Reprendre la question ci-dessus en utilisant une des techniques de validation croisée *«5-Fold CV»* ou *«10-Fold CV»* de votre choix. Justifier brièvement votre choix.
3. 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 250 points (2 couleurs de votre choix) et ajouter au graphique les courbes (trois en tout, similaires à celles de la figure 5.7 page 185 dans ISL) 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) ;
- l'analyse discriminante retenue (LDA ou QDA).

QUESTION N° 2 (5 points). On dispose d'un échantillon de 48 observations obtenues lors de prélèvements par carottage dans un contexte d'exploration pétrolière (voir données `rock`, disponibles dans R). Les données sont de la forme $\{\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i3})^\top, i = 1, \ldots, 48\}$. On considère que chaque $\mathbf{x}_i, i = 1, \ldots, 48$ est une observation d'un vecteur aléatoire $X = (X_1, X_2, X_3, X_4)^\top$ dont la distribution de probabilité exacte est inconnue. Considérer $X_1 = \text{area}, X_2 = \text{peri}, X_3 = \text{shape}, X_3 = \text{perm}$.

On s'intéresse à l'estimation du paramètre θ (la corrélation) défini par

$$\theta = \texttt{corr}\left(\log|X_1 - 3X_2|,\ \max\{250X_3, X_4\}\right).$$

En utilisant la technique de ré-échantillonnage *«Bootstrap»* (fonction `boot()`) avec 2000 répétitions :

a) 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) Déduire des résultats qui précèdent un intervalle de confiance pour θ au niveau de confiance 95%. Commenter brièvement.

QUESTION N° 3 (5 points). Exercice 7 page 262 ISL (An Introduction to Statistical Learning)