DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL

MTH6312 - MÉTHODES STATISTIQUES D'APPRENTISSAGE

Devoir nº 4 - Automne 2018

Date de remise : 27 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 (15 points). On considère de nouveau la base de données `Carseats` (Question 2 du devoir nº 2) dont la description est disponible dans le *package* ISLR.

Dans cette question, on recherche un bon modèle de régression afin de prédire `Sales` (les ventes) en utilisant les autres variables de la base de données comme *inputs*.

Précisions. Considérer le k-fold CV (avec $k = 5$ ou $k = 10$) comme méthode de validation croisée à utiliser dans les questions qui suivent.

a) **Séparation des données.**

Diviser les données en deux groupes qui serviront pour toutes les questions qui suivent.
Pour cela, utiliser la fonction `sample()` de R afin de choisir un échantillon (sans remise) de taille 200 pour les données d'entraînement, et le reste des données (200) constituera les données de test.

Remarque. Un germe (seed) est nécessaire pour que la fonction `sample()` produise toujours les mêmes résultats (voir les directives sur la première page du devoir).

b) **Approche par meilleurs sous ensembles.**

1. Avec les données d'entraînement, utiliser la fonction `regsubset()` de R pour déterminer les meilleurs modèles linéaires de ℓ variables (selon R^2 ajusté), notés $M_\ell, \ell = 0, \ldots, 11$. Déterminer ensuite le meilleur des 12 modèles par validation croisée.
2. Avec les données de test, évaluer l'erreur de test (i.e. la moyenne des carrés des erreurs de prévision) du meilleur modèle retenu.

c) **Approche Ridge.**

1. Avec les données d'entraînement, ajuster un modèle de régression ridge (fonction `glmnet()` de R) et produire un graphique similaire à celui de la partie gauche de la figure 6.4 page 216 dans ISL. Déterminer ensuite la valeur appropriée de λ (modèle optimal) par validation croisée.
2. Avec les données de test, évaluer l'erreur de test (i.e. la moyenne des carrés des erreurs de prévision) du modèle optimal de la méthode ridge.

d) **Approche Lasso.**

1. Avec les données d'entraînement, ajuster un modèle de régression lasso (fonction `glmnet()` de R) et produire un graphique similaire à celui de la partie gauche de la figure 6.6 page 220 dans ISL. Déterminer ensuite la valeur appropriée de λ (modèle optimal) par validation croisée.
2. Avec les données de test, évaluer l'erreur de test (i.e. la moyenne des carrés des erreurs de prévision) du modèle optimal de la méthode ridge.

e) **Approche composantes principales**.

1. Avec les données d'entraînement, utiliser la fonction `pcr()` de R pour déterminer les modèles linéaires avec M composantes principales. Déterminer ensuite la valeur optimale de M (meilleur modèle) par validation croisée.
2. Avec les données de test, évaluer l'erreur de test (i.e. la moyenne des carrés des erreurs de prévision) du meilleur modèle retenu.

f) **Approche moindres carrés partiels**.

1. Avec les données d'entraînement, utiliser la fonction `plsr()` de R pour ajuster les modèles linéaires de M composantes (directions). Déterminer ensuite la valeur optimale de M (meilleur modèle) par validation croisée.
2. Avec les données de test, évaluer l'erreur de test (i.e. la moyenne des carrés des erreurs de prévision) du meilleur modèle retenu.

g) **Commentaires et conclusion**. Commenter sur les résultats obtenus. Y a-t-il une différence importante entre les taux d'erreur des 5 approches (ou méthodes d'apprentissage) ? Quelle modèle utiliseriez-vous pour calculer les prévisions des ventes ? Commenter brièvement.

QUESTION N^{o} 2 (5 points). Exercice 6 page 299 ISL

Remarque. Les données `Wage` sont disponibles dans le *package* ISLR.