DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL

MTH6312 - MÉTHODES STATISTIQUES D'APPRENTISSAGE

Devoir nº 2 - Automne 2018

Date de remise : 15 octobre avant 23h55 (en pdf dans Moodle)

DIRECTIVES :

- ✓ Inclure dans votre rapport le code R que vous avez utilisé.
- ✓ 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) On considère les données `Carseats` dont la description est disponible dans le package ISLR (voir site du cours).

a) En utilisant la fonction `lm()` de R, effectuer l'ajustement d'un modèle de régression linéaire simple avec `Sales` (les ventes) comme variable dépendante (output Y) et `Price` (le prix) comme variable indépendante (input X_1).
Produire les résultats avec la fonction `summary()`, les graphiques diagnostiques des résidus avec la fonction `plot()` et commenter brièvement sur le points suivants :

1. Le modèle linéaire simple explique-t-il bien le lien entre l'output Y et l'input X_1 ?
2. Quelle est la valeur de l'output Y prédite par le modèle pour un input $X_1 = 117{,}5$?
   Donner un intervalle de confiance et un intervalle de prévision pour cette valeur au niveau de confiance 95%.
3. Les graphiques des résidus indiquent-ils une anomalie pour l'ajustement du modèle ? Y a-t-il présence de points influents.

b) Tracer le nuage de points (en bleu) et ajouter au graphique les éléments suivants :

1. la droite de régression (en rouge) obtenue en a) ;
2. en utilisant la fonction `seq()` de R, subdiviser en 300 valeurs l'intervalle allant du minimum $\min(X_1)$ au maximum $\max(X_1)$ des valeurs de X_1. Utiliser ces valeurs pour construire deux courbes représentant les limites de confiance supérieures et inférieures pour la fonction de régression (i.e. $\beta_0 + \beta_1 X_1$) à 95% ;
3. En utilisant les 300 valeurs de X_1 de la sous question précédente, construire deux courbes représentant les limites de prévision supérieures et inférieures pour Y à 95%.

Remarque : Les courbes des limites de confiance et celles de prévision doivent être de couleurs et de motifs différents. Pour cela, utiliser les options «`col`» et «`lty`» dans les fonctions graphiques de R telle que `plot()`.

c) On considère la variable US (le magasin est au USA ou ailleurs) comme deuxième input (X_2) et un modèle de régression linéaire avec interaction, i.e. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$. Procéder à l'ajustement de ce modèle et, à l'aide d'un seul test de seuil critique 5%, dire si les ventes (`Sales`) en fonction du prix (`Price`) peuvent être représentées par une seule équation, que le magasin soit aux USA ou non.

d) On considère à présent un modèle de régression avec Y comme output, et toutes les autres variables quantitatives comme input (7 variables). Procéder à l'ajustement de ce modèle.

1. identifier les variables qui ne contribuent pas significativement au modèle ;
2. obtenir les graphiques diagnostiques des résidus et dire s'il y a présence d'anomalies, de points influents, etc ;
3. proposer un modèle simplifié et justifier votre choix.

QUESTION N° 2 (10 points). Des mesures de deux variables (X_1, X_2) sont obtenues sur un équipement industriel utilisé de façon continue, ainsi que l'état Y (présence ou absence d'anomalies) de l'équipement. On dispose d'un échantillon aléatoire de 250 de ces mesures (voir fichier *Equipement.csv* sur le site du cours), 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.

Dans cette question les 170 premières données constituent les *données d'entraînement* et les 80 dernières sont les *données de test*. Trois méthodes de classification (régression linéaire, KNN, régression logistique) sont envisagées.

a) Ajuster le modèle de régression linéaire (contexte de classification) avec les données d'entraînement. Utiliser le résultat pour classifier les données de test et déterminer le taux d'erreur.

b) Considérer le classificateur du KNN (avec la distance euclidienne). Pour chaque valeur du nombre de voisins K (utiliser au moins 50 valeurs différentes de K) : entraîner le classificateur du KNN sur les données d'entraînement, classifier les données de test et déterminer le taux d'erreur. Tracer ensuite la courbe du taux d'erreur en fonction de K et déterminer la valeur optimale de K.

c) Ajuster le modèle de régression logistique (contexte de classification) aux les données d'entraînement. Utiliser le résultat pour classifier les données de test et déterminer le taux d'erreur.

d) Produire un seul graphique, similaire à ceux des figures 2.1 à 2.3 pages 13 à 16 de ESL, contenant les trois courbes délimitant les deux classes selon chacune des méthodes de classification : la régression linéaire, la régression logistique, la méthode du KNN avec la valeur optimale de K obtenue en b).

e) Supposons que l'on dispose d'une nouvelle donnée : $X_1 = 9{,}5$; $X_2 = 13,5$. En utilisant l'ensemble des 250 observations et chacune des trois méthodes, quelle prévision peut-on faire sur l'état réel de l'équipement ? Commenter brièvement.