Université de Thiès **Année Académique 2019-2020**
U.F.R SES **Master 1 Sciences de données**

Apprentissage Statistique

Travail individuel.
à rendre avant 31 janvier 2021 à $23h59$

Exercice 1 On considère le modèle de régression $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \;\; 1 \leq i \leq n$, que l'on écrit sous la forme $Y = X\beta + \varepsilon$. Les $x_{i,j}$ sont des variables exogènes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. On a observé :

$$X^{'}X = \begin{pmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{pmatrix}, \quad X^{'}Y = \begin{pmatrix} 15 \\ 20 \\ 10 \end{pmatrix}, \quad Y^{'}Y = 59.5.$$

1. Déterminer n, la moyenne des $x_{i,2}$, le coefficient de corrélation des $x_{i,1}$ et des $x_{i,2}$.
2. Estimer $\beta_0, \beta_1, \beta_2, \sigma^2$ par la méthode des moindres carrés ordinaires.
3. Calculer pour β_1 un intervalle de confiance à 95% et tester $\beta = 0.8$ a niveau 10%
4. Tester $\beta_0 + \beta_1 = 3$ contre $\beta_0 + \beta_1 \neq 3$ au niveau 5%
5. Calculer y et déduire le coefficient de détermination ajusté R^2
6. Construire un intervalle de prévision à 95% de y_{n+1} si $x_{n+1,1} = 3$ et $x_{n+1,2} = 0.5$.

Exercice 2 Soient $X \in \mathbb{R}^{n \times (p+1)}$ la matrice contenant les données dont la $i^{\text{ème}}$ ligne est $(\boldsymbol{x}_i^{'}, 1)$ avec $\boldsymbol{x}_i^{'} = (x_1, \ldots, x_p)$ et $Y \in \mathbb{R}^n$ vecteur contenant les étiquettes y_i. L'estimateur des moindre carrés le vecteur

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T X)^{-1} X^T Y = \min_{\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - (\langle \alpha, \boldsymbol{x}_i \rangle + \beta))^2$$

1. Programmez une fonction *regression*(X, Y) qui renvoie l'estimateur des moindre carrés.
   Utiliser votre fonction de régression sur le jeu de données Boston House Prices (à charger avec la fonction *datasets.load boston*(). Comparez les vecteurs $\hat{\alpha}$ et $\hat{\beta}$ renvoyés par votre fonction avec les attributs *coef_* et *intercept_* d'un régresseur de type *linear_model.LinearRegression*.
   Quelques fonctions utiles : *dot*(), *transpose*(), *pinv*().
2. Écrire une fonction *regress*(X, α, β) qui renvoie le vecteur $\hat{Y}$ des étiquettes prédites tel que $\hat{y}_i = \langle \alpha, \boldsymbol{x}_i \rangle + \beta$
3. Calculer $\hat{\epsilon} = \|Y - \hat{Y}\|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ l'erreur au sens des moindres carrés du régresseur appris sur l'ensemble du jeu de données Boston.
4. Dans certains cas, la matrice $X^T X$ n'est pas inversible. Pour remédier à ce problème, on ajoute un ridge λI_{p+1} à cette matrice où I_{p+1} est la matrice identité d'ordre $p+1$.

Cela correspond à une légère modification du problème d'optimisation qui pénalise la taille des coefficients. Le vecteur des moindre carrés généralisés est donné par :

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = (X^T X + \lambda I_{p+1})^{-1} X^T Y = \min_{\alpha\in\mathbb{R}^p,\beta\in\mathbb{R}} \sum_{i=1}^{n} (y_i - (\langle \alpha, \boldsymbol{x}_i\rangle + \beta))^2 + \lambda\|\alpha\|_2^2$$

(a) Programmez une fonction *ridge_regression*(X, Y, lambda) qui renvoie l'estimateur des moindre carrés généralisés. Comparez à nouveau les vecteurs $\hat{\alpha}$ et $\hat{\beta}$ obtenus pour le paramètre lambda = 1 sur le jeu de données Boston avec les attributs *coef_* et *intercept_* d'un régresseur de type *linear_model.Ridge*

(b) Tracez l'évolution des coefficients du vecteur $\hat{\alpha}$ en fonction du paramètre de régularisation lambda pour des valeurs entre 0.001 et 1000. Quelles variables semblent le mieux expliquer le prix des maisons à Boston ?

(c) Trouvez par un moyen approprié la meilleure valeur pour le paramètre lambda. Apprenez ensuite un régresseur avec cette valeur sur l'ensemble du jeu de données Boston et calculez l'erreur au sens des moindres carrés sur ce même échantillon.

5. La formulation Lasso est une variante de la régression linéaire régularisée. La pénalisation du vecteur des coefficients se fait ici avec la norme $||_1$ a la place de la norme euclidienne $||_2$. Soit $\alpha \in \mathbb{R}^p$, $\|\alpha\|_1 = \sum_{i=1}^{p} |\alpha_i|$. Il s'ensuit des solutions dites parcimonieuses, c'est-à-dire que de nombreux coefficients sont égaux à zéro. Le problème d'optimisation s'écrit alors :

$$\min_{\alpha\in\mathbb{R}^p,\beta\in\mathbb{R}} \sum_{i=1}^{n} (y_i - (\langle \alpha, \boldsymbol{x}_i\rangle + \beta))^2 + \lambda\|\alpha\|_1$$

(a) En utilisant la *classe linear_model.Lasso*, tracez l'évolution des coefficients du vecteur $\hat{\alpha}$ en fonction de la valeur du paramètre *lambda*. Quelles variables semblent le mieux expliquer le prix des maisons à Boston ? Sont-elles les mêmes que celles trouvées à l'exercice précédent ? Comment se comportent les autres variables lorsque la valeur de *lambda* augmente ?

(b) Trouvez par un moyen approprié la meilleure valeur pour le paramètre lambda. Apprenez ensuite un régresseur avec cette valeur sur l'ensemble du jeu de données Boston et calculez l'erreur au sens des moindres carrés sur ce même échantillon.