| ACT6100 | Analyse de données |
|---|---|
| H2019 | Série 3 |

**Apprentissage supervisé - Introduction**

1. La base de données *dataEXO1.csv* disponible sur le site du cours contient 100 observations d'une variable réponse $Y$ et d'une variable explicative $X$.
   (a) Réaliser un nuage de points représentant ces données. En R, estimer les paramètres d'un modèle de régression linéaire et ajouter ce modèle sur le graphique. Est-ce que les paramètres sont significatifs ? Quelle est l'erreur quadratique moyenne ?
   (b) En R, ajuster maintenant un modèle quadratique et ajouter ce modèle sur le graphique. Est-ce que les paramètres sont significatifs ? Quelle est l'erreur quadratique moyenne ?
   (c) Si on utilise l'algorithme des $K$ plus proches voisins, quelle valeur de $K$ minimise l'erreur quadratique d'entrainement ? Ajouter ce modèle au graphique. Quelle est la valeur de l'erreur quadratique moyenne d'entrainement du modèle optimal ?
   (d) Si on utilise l'algorithme des $K$ plus proches voisins avec validation croisée (*Leave one out*), quelle valeur de $K$ minimise l'erreur quadratique de validation ? Ajouter ce modèle au graphique. Quelle est la valeur de l'erreur quadratique moyenne de validation du modèle optimal ?
   (e) Pour une nouvelle observation $(x^*, y^*) = (333.2522, 99\,508.44)$, utiliser les 4 modèles précédents afin d'obtenir une prédiction et calculer, à chaque fois, l'erreur quadratique de prédiction.
2. On considère une petite base de données contenant les valeurs présentées à la Table 1. Répondre aux questions ci-dessous « à la main ».

| $i$ | $Y_i$ | $X_i$ |
|---|---|---|
| 1 | 4 | 1 |
| 2 | 2 | 4 |
| 3 | 8 | 2 |
| 4 | 5 | 9 |
| 5 | 3 | 7 |

Table 1 – Base de données.

   (a) Ajuster un modèle de régression linéaire en minimisant l'erreur quadratique moyenne d'entrainement. Écrire l'équation du modèle.
   (b) Pour un modèle des 2 plus proches voisins, ajuster le modèle sans validation croisée et écrire l'équation du modèle.
   (c) Pour un modèle des 2 plus proches voisins, calculer la valeur du vMSE avec *leave one out cross validation*.
3. La base de données *swmotorcycle* disponible dans la librarie *CASdatasets* contient des fréquences de sinistre observées (variable **ClaimNb**) pour 64 548 assurés ainsi que l'âge de la personne assurée (variable **OwnerAge**) et l'exposition (variable **Exposure**), en années.
   (a) En considérant uniquement les trois variables mentionnées, réaliser un « nettoyage » de la base de données. Justifier.

(b) Afin de modéliser la fréquence des sinistres ($N$), on souhaite utiliser un modèle Poisson avec

$$E[N] = (\mathbf{Exposure}) \exp\left(\beta_0 + \beta_1 \mathbf{OwnerAge}\right).$$

Calculer $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

(c) Généralement, dans l'idée de construire une table de tarification, on divise la variable **OwnerAge** en groupes. On peut, par exemple, construire un modèle avec deux classes telles que $x \in C_1$ si l'âge de l'assuré est plus petit ou égal à $K$ années et $x \in C_2$ sinon. Déterminer, **(i)** sans validation croisée et **(ii)** avec une 12-validation croisée (groupes de tailles égales), le modèle optimal en vous basant sur l'erreur quadratique moyenne. Quelles seront les fréquences moyennes par groupe pour un assuré dont l'exposition est unitaire ?

**Réponses**

1. (a) Les deux paramètres sont significatifs. Le MSE est 543 910 771 909.
   (b) Le paramètre $X$ n'est pas significatif. Un modèle avec une ordonnée à l'origine et une variable $X^2$ conduit à un MSE de 434 064 644 839.
   (c) $K = 1$ avec un MSE de 0.
   (d) $K = 10$ avec un MSE de 572 388 692 923.
   (e) On obtient, dans l'ordre, 6 609 874 222, 2 580 396 252, 17 333 311 211, 2 338 794 222.
2. (a) $\widehat{f}(X) = 5.1327 - 0.159292X$
   (b)
$$\widehat{f}(X) = \begin{cases} 6, & X \leq 2.5 \\ 5, & 2.5 < X \leq 4.5 \\ 2.5, & 4.5 < X \leq 6.5 \\ 4, & X > 6.5 \end{cases}$$
   (c) 48.5
3. (a) –
   (b) $\widehat{\beta}_0 = -2.13447$ et $\widehat{\beta}_1 = -0.06035$
   (c) **(i)** division à 77 ans $\rightarrow$ 0.01064113 et $1.300423 \times 10^{-7}$ **(ii)** division à 30 ans $\rightarrow$ 0.03132543 et 0.006331919