Université Paris-Saclay : faculté de médecine
Master : santé publique
Introduction aux data sciences

Examen + corrigé du 31 août 2022 à 14h. Durée : 1h30

I. Apprentissage statistique : notions générales (10 points)

A. Le tableau suivant fournit un jeu de données synthétique d'apprentissage contenant six observations, trois variables explicatives et une variable réponse binaire (ou catégorielle). On souhaite ajuster un modèle des k plus proches voisins (knn) sur ce jeu de données pour

Obs.	X_1	X_2	X_3	distance à $(0,0,0)$	Y
1	0	3	0	3	rouge
2	2	0	0	2	rouge
3	0	1	3	$\sqrt{10} \approx 3.2$	rouge
4	0	1	2	$\sqrt{5} \approx 2.2$	vert
5	-1	0	1	$\sqrt{2} \approx 1.4$	vert
6	1	1	1	$\sqrt{3} \approx 1.7$	rouge

prédire la réponse d'un point test $X_1 = X_2 = X_3 = 0$.

a. Calculer la distance euclidienne entre chaque point du jeu de données d'apprentissage et le point test $X_1 = X_2 = X_3 = 0$. [1]

corrigé : *voir tableau. (1.5 points)*

b. Quel est la prédiction de la réponse du point test lorsque $k = 1$.

corrigé : *le plus proche voisin est vert. (1 point)*

c. Quel est la prédiction de la réponse du point test lorsque $k = 3$.

corrigé : *rouge, les trois plus proches voisins sont un vert et deux rouges. (1 point)*

d. Supposons que la frontière de classement du modèle réel (inconnue en pratique) est hautement non linéaire. Devrions-nous nous attendre à ce que la meilleure valeur pour k soit grande ou petite ? Justifier ?

corrigé : *Petit. Un petit k serait flexible pour une frontière de décision non-linéaire, alors qu'un grand k essaierait de s'adapter à une frontière plus linéaire parce qu'il prend en compte plus de points. (1.5 point)*

1. Rappelons que la distance euclidienne entre deux vecteurs $u = (u_1, u_2, \ldots, u_p)$ et $v = (v_1, v_2, \ldots, v_p)$ est donnée par $\|u - v\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_p - v_p)^2}$.

B. Pour chacune des situations suivantes, indiquer s'il est préférable d'ajuster un modèle complexe ou un modèle simple. Justifier intuitivement votre réponse.

a. Un jeu de données avec un grand nombre d'observations n et un petit nombre de variables explicatives p.

corrigé : *Un modèle complexe et donc flexible est préférable à un modèle simple. La grande taille d'échantillon permettra l'estimation de la multitude de paramètres. (1 point)*

b. Un jeu de données avec un petit nombre de d'observations n et un très grand nombre de variables explicatives p.

corrigé : *Un modèle simple est préférable. Risque de sur-apprentissage d'un modèle complexe. (1 point)*

c. La relation entre les variables explicatives et la variable réponse est hautement non-linéaire.

corrigé : *Un modèle complexe (grand degré de liberté) permet un ajustement aux données. (1 point)*

d. La variance du terme d'erreur du modèle *i.e.* $\text{Var}(\varepsilon) = \sigma^2$ est très élevée où

$$Y = f(X) + \varepsilon.$$

corrigé : *Un modèle simple est préférable. Un modèle complexe ajustera le bruit (risque de forte variance des paramètres). (1 point)*

II. Régression linéaire : exploitation des paramètres (6.5 points)

Supposons que nous avons un jeu de données avec 5 variables explicatives :

- X_1 correspond à la moyenne GPA (*Grade Point Average*).
- X_2 correspond au Quotient Intellectuel.
- X_3 correspond au sexe (1 pour femme et 0 pour homme).
- X_4 correspond à l'interaction entre la moyenne GPA et le QI.
- X_5 correspond à l'interaction entre la moyenne GPA et le sexe.
- Y correspond au salaire de départ après l'obtention du diplôme (en milliers de dollars).

Supposons que nous avons ajusté un modèle de régression linéaire multiple. Les coefficients de régression obtenu par minimisation des moindres carrés ont pour valeurs $\widehat{\beta}_0 = 50$, $\widehat{\beta}_1 = 20$, $\widehat{\beta}_2 = 0.07$, $\widehat{\beta}_3 = 35$, $\widehat{\beta}_4 = 0.01$ et $\widehat{\beta}_5 = -10$.

Le modèle est donné par : $Y = 50 + 20gpa + 0.07qi + 35sexe + 0.01(gpa \times qi) - 10(gpa \times sexe)$. *Nous avons* $Y = 50 + 20x_1 + 0.07x_2 + 35sexe + 0.01(x_1 \times x_2) - 10(x_1 \times sexe)$. *Donc homme (*$sexe = 0$*)* $y = 50 + 20x_1 + 0.07x_2 + 0.01(x_1 \times x_2)$ *et femme (*$sexe = 1$*)* $Y = 50 + 20x_1 + 0.07x_2 + 35 + 0.01(x_1 \times x_2) - 10(x_1)$. *Donc affirmation vraie dès que le score GPA est suffisamment élevé.*

A. Parmi les affirmations suivantes, lesquelles sont vraies ? Justifier.

a. Pour des valeurs fixées de QI et de GPA, les hommes gagnent en moyenne plus que les femmes.

corrigé : *Non ! (0.5 point)*

b. Pour des valeurs fixées de QI et de GPA, les femmes gagnent en moyenne plus que les hommes.

corrigé : *Non ! (0.5 point)*

c. Pour des valeurs fixées de QI et de GPA, les hommes gagnent en moyenne plus que les femmes à condition que la valeur de GPA soit suffisamment élevée.

corrigé : *Oui ! (2 points)*

d. Pour des valeurs fixées de QI et de GPA, les femmes gagnent en moyenne plus que les hommes à condition que la valeur de GPA soit suffisamment élevée.

corrigé : *Non ! (0.5 point)*

B. Prédire le salaire d'une femme ayant un QI de 110 et une moyenne GPA de 4.0.
corrigé : $Y(sexe = 1, qi = 110, GPA = 4.0) = 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01(4 * 110) - 10 * 4 = 137.1$ *! (2 point)*

C. Répondre par vrai ou faux à l'affirmation suivante : *le coefficient de régression de la variable d'interaction GPA / IQ est très petit, il y a très peu de preuves d'existence d'un effet d'interaction entre les deux variables.* Justifier.

corrigé : *Faux. Il faut faire appel à un test pour examiner si la valeur du paramètre est significativement différente de zéro. (1 point)*

III. La grande famille des modèles de régression (3.5 points)

A. Classer les modèles suivants en une famille de modèles et préciser les sous-familles. Préciser l'ensemble des valeurs des hyper-paramètres qui permettent de balayer chaque famille de modèles. Préciser la valeur de l'hyper-paramètre qui permet de restreindre une famille de modèles à une sous-famille.

1. Régression linéaire multiple
2. Régression elasticnet
3. Régression ridge
4. Régression lasso

corrigé : *elasticnet* ($\alpha > 0$ *et* $\lambda > 0$) $\longrightarrow$ *lasso* ($\alpha = 0$ *et* $\lambda > 0$) $\longrightarrow$ *régression linéaire* ($\lambda = 0$)