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

Examen + corrigé du 12 juin 2019. Durée : 1h30

I. Apprentissage statistique : notions générales (11 points)

A. Expliquer en quelques mots la(les) différence(s) entre un modèle de classification par les k plus proches voisins et un modèle de régression par les k plus proches voisins.

corrigé : *Y est quantitatif pour la régression et binaire pour la classification. Les deux méthodes font appel aux données réponse y des voisins d'une observation x. Le régression prédit la réponse de x par la moyenne des réponses des voisins. La classification prédit la réponse de x à l'aide d'un vote majoritaire simple des réponses des voisins. (2 points)*

B. 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)*

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 $\|u - v\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_p - v_p)^2}$.

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)*

C. 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. Régression pénalisée (2 points)

Supposons qu'on estime des coefficients de régression d'un modèle de régression linéaire pénalisée en minimisant les deux expressions suivantes

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{1}$$

et

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \tag{2}$$

Le tableau suivant résume un ensemble de quantités qui peuvent être étudiées pour les deux modèles précédents :

Modèle	**RSS** d'apprentissage	**RSS** de test	Variance des coefficients	Carré du biais des coefficients
(1)	4	2	3	4
(2)	4	2	3	4

Pour chaque quantité, identifier la vraie affirmation lorsque $\lambda \to 0$

corrigé : *voir le tableau (2 points)*

1. Augmente initialement, ensuite, elle diminue en formant un U inversé.
2. Diminue initialement, ensuite, elle augmente en formant U.
3. Ne fait qu'augmenter.
4. Ne fait que diminuer.
5. Reste constante.

IV. Classification (5 points)

A. Supposons qu'on s'intéresse à un jeu de données d'un groupe d'étudiants d'un cours d'apprentissage statistique en M1 composé de 2 variables explicatives et une variable réponse : X_1 nombre d'heures étudiées, X_2 la moyenne *Grade Point Average* en premier cycle à l'université et Y correspond à l'obtention de la mention "A" (versus, "C", "D" et "F"). L'estimation des coefficients de régression logistique nous donne : $\widehat{\beta}_0 = -6$, $\widehat{\beta}_1 = 0.05$ et $\widehat{\beta}_2 = 1$.

1. Estimer la probabilité qu'un étudiant ayant consacré 40h d'étude et une moyenne GPA 3.5 en premier cycle, obtienne une mention "A".

   corrigé : 37.75% *(1.5 points)*

2. Quelle est le temps d'étude nécessaire pour que cet étudiant ait 50% de chances d'obtenir une mention "A".

corrigé : *50 heures (1.5 points)*

B. Supposons que nous avons des jeux de données apprentissage et test de tailles égales et deux méthodes de classification en compétition. La première méthode est une régression logistique avec une erreur d'apprentissage de 0.2 et une erreur de test de 0.3. La seconde méthode est une méthode des k plus proches voisins avec $k = 1$ avec une erreur moyenne (calculée sur l'ensemble du jeu de données apprentissage et test) de 0.18. Au vu de ces résultats, quelle est la méthode à retenir pour classer une nouvelle observation ? Justifier.

corrigé : *Pour k-nn avec $k = 1$, le taux d'erreur d'apprentissage est de 0 car pour toute observation d'apprentissage, le plus proche voisin sera elle même. k-nn a donc un taux d'erreur de test de 0.36. Donc la régression logistique réalise une erreur de test meilleure (0.3). (2 points)*

V. Fléau de la dimension (3 points)

Lorsque le nombre de variables explicatives p est grand, les performances d'un modèle des k plus proches voisins se dégradent. Rappelons que cette méthode utilise les observations proches de l'observation test pour laquelle une prédiction doit être effectuée. Ce phénomène est appelé **fléau de la dimension** (*curse of dimensionality* en anglais). Nous allons maintenant essayer d'illustrer ce phénomène.

1. Supposons que nous avons un jeu de données. Chaque donnée est un couple (X, Y) où X est en dimension $p = 1$. Nous supposons que X est uniformément distribué sur l'intervalle $[0, 1]$. Supposons que nous souhaitons prédire la réponse d'une observation test en utilisant uniquement des observations situées à moins de 10% de l'intervalle de valeurs de X. Par exemple, pour prédire la réponse d'une observation test avec $x = 0.6$, nous utiliserons des observations comprises dans l'intervalle $[0.55, 0.65]$. En moyenne, quelle fraction des observations allons-nous utiliser pour calculer la prédiction d'une observation test [2] ?

   corrigé : *En moyenne 10% (la proportion du segment en question !) (1 point)*

2. Supposons maintenant que $X = (X_1, X_2)$ est de dimension $p = 2$. Nous supposons que le couple (X_1, X_2) est uniformément réparti sur le carré $[0, 1] \times [0, 1]$. Nous souhaitons prédire la réponse d'une observation d'une donnée test à l'aide des observations situées dans la limite des 10% de l'intervalle des valeurs de X_1 et de 10% de l'intervalle des valeurs de X_2 plus proche de l'observation de test. Par exemple, pour prédire la réponse pour une observation test avec $x_1 = 0.6$ et $x_2 = 0.35$, nous utiliserons les observations comprises dans l'intervalle $[0.55, 0.65]$ pour X_1 et dans l'intervalle $[0.3, 0.4]$ pour X_2. En moyenne, quelle fraction des observations allons-nous utiliser pour calculer la prédiction d'une observation test ?

   corrigé : *En moyenne 1% (la proportion du volume en question !) (1 point)*

3. Conclure !

   corrigé : *En moyenne, une proportion de 10^{-p} en dimension p (1 point)*

2. Pour simplifier, ignorer les cas où $X < 0.05$ et $X > 0.95$.