QCM **Grande dimension**

Test
Examen du 16/11/2022

Instructions :

- Le sujet comprend 19 questions. Les questions faisant apparaître le symbole ♣ peuvent présenter plusieurs bonnes réponses. Les autres ont une unique bonne réponse. Des points négatifs pourront être affectés à de *mauvaises* réponses.
- Seul le questionnaire de la page 5 est à rendre. Vous commencerez par renseigner votre nom et prénom dans la case prévue ainsi que le numéro d'étudiant.
- Il faut **colorier** les cases correspondants aux bonnes réponses (sur la page 5), mettre une croix dans la case n'est **pas suffisant**. Les cases devront être **coloriées avec un stylo noir** (pas de crayon papier, de stabilo...).
- Le barème sera effectué de la façon suivante :
  - Aucune case coloriée entrainera une note de 0 sur la question.
  - Pour les questions à une seule bonne réponse (sans le symbole ♣), un nombre de points sera affecté (par exemple +2) si la bonne case est cochée. Un nombre de points sera retranché (par exemple -1) si une mauvaise case est coloriée ou si plusieurs cases sont coloriées.
  - Pour les questions avec plusieurs bonnes réponses (avec le symbole ♣), un nombre de points (par exemple +0.5) sera affecté pour chaque bonne réponse coloriée et pour chaque mauvaise réponse non coloriée. Un nombre de points (par exemple -0.5) sera retranché pour chaque mauvaise réponse coloriée et pour chaque bonne réponse non coloriée.
- La correction étant automatique, un non respect des consignes aura forcément un impact sur la note finale.

Durée : 1 heure.

On se placera dans tout le devoir dans un modèle de régression : on dispose de n observations i.i.d. $(x_i, y_i), i = 1, \dots, n$ où x_i est à valeurs dans $\mathbb{R}^p$ et y_i dans $\mathbb{R}$ avec

$$y_i = m(x_i) + \varepsilon_i = m(x_{i1}, \dots, x_{ip}) + \varepsilon_i.$$

Les termes d'erreur ε_i sont i.i.d.

Question 1 On suppose que les variables explicatives $X_j, j = 1, \dots, p$ ne sont pas à la même échelle et on considère l'estimateur des MCO du modèle linéaire. Cocher la (ou les) assertion(s) vraie(s) :

■ L'algorithme des MCO n'est pas sensible à la réduction des variables explicatives

B Il est très important de réduire les variables explicatives avant de calculer les estimateurs des MCO

C aucune réponse n'est correcte

Question 2 On suppose que les variables explicatives $X_j, j = 1, \dots, p$ ne sont pas à la même échelle et on considère la distance euclidienne dans $\mathbb{R}^p$ pour calculer l'estimateur des k plus proches voisins. Cocher la (ou les) assertion(s) vraie(s) :

A L'algorithme des plus proches voisins n'est pas sensible à la réduction des données

■ Il est préférable de réduire les variables explicatives pour calculer les plus proches voisins d'une nouvelle observation

C aucune réponse n'est correcte

Question 3 ♣ Lorsque $p > n$

- [A] l'estimateur des MCO n'existe pas
- [■] la matrice $\mathbb{X}^t\mathbb{X}$ n'est pas inversible
- [C] l'estimateur des MCO existe et est unique
- [■] l'estimateur des MCO existe et n'est pas unique
- [E] *Aucune de ces réponses n'est correcte.*

Question 4 ♣ Soit k un entier plus petit que n. On désigne par $\widehat{m}_k$ l'estimateur des k plus proches voisins. Cocher la (ou les) assertion(s) vraie(s) :

- [■] $\widehat{m}_k$ est un estimateur non-paramétrique
- [■] $\widehat{m}_k$ possède généralement un biais faible pour de petites valeurs de k
- [C] $\widehat{m}_k$ est un estimateur paramétrique
- [■] $\widehat{m}_k$ possède généralement une variance faible pour de grandes valeurs de k
- [E] $\widehat{m}_k$ aura tendance à surajuster si k est trop grand
- [F] $\widehat{m}_k$ possède généralement un biais élevé pour de petites valeurs de k
- [■] $\widehat{m}_k$ aura tendance à surajuster si k est trop petit
- [H] $\widehat{m}_k$ possède généralement une variance élevée pour de grandes valeurs de k
- [I] *Aucune de ces réponses n'est correcte.*

Question 5 ♣
Soit $\widehat{m}_h$ un estimateur à noyau dont

- le carré du biais est de l'ordre de $C_1 h^4$;
- la variance est de l'ordre $C_2/(nh^p)$,

où C_1 et C_2 sont des constantes qui dépendent uniquement de la fonction m. On appelle fenêtre optimale la valeur de h qui minimise l'erreur quadratique (carré du biais + variance) et vitesse optimale l'erreur quadratique qui correspond à la fenêtre optimale. Cocher la (ou les) assertion(s) vraies (C_3 et C_4 désignent des constantes qui dépendent de C_1 et C_2) :

- [■] La vitesse optimale est de l'ordre de $C_4 n^{-4/(p+4)}$
- [B] La vitesse optimale est de l'ordre de $C_4 n^{-2/(p+2)}$
- [C] La fenêtre optimale est de l'ordre de $C_3 n^{-1/(p+1)}$
- [D] La vitesse optimale est de l'ordre de $C_4 n^{2/(p+1)}$
- [E] La fenêtre optimale est de l'ordre de $C_3 n^{-1/(p+2)}$
- [F] La fenêtre optimale est de l'ordre de $C_3 n^{1/(p+1)}$
- [G] La vitesse optimale est de l'ordre de $C_4 n^{2/p}$
- [H] La vitesse optimale est de l'ordre de $C_4 n^{-2/(p+1)}$
- [■] La fenêtre optimale est de l'ordre de $C_3 n^{-1/(p+4)}$
- [J] La fenêtre optimale est de l'ordre de $C_3 n^{1/p}$
- [K] *Aucune de ces réponses n'est correcte.*

Question 6 ♣ Les assertions suivantes sont liées au fléau de la dimension. Cocher la (ou les) assertion(s) vraies :

- [A] Les estimateurs paramétriques sont toujours efficaces lorsque p est grand
- [B] L'erreur quadratique des estimateurs non paramétrique tend vers 0 de plus en plus vite lorsque p augmente
- [■] Les estimateurs non paramétriques sont généralement peu efficaces lorsque p est grand
- [■] Il est difficile de trouver des observations proches du point où on cherche à estimer la fonction de régression lorsque p est grand
- [E] Les estimateurs paramétriques sont toujours meilleurs que les estimateurs non paramétriques
- [F] Les estimateurs non paramétriques sont peu efficaces en classification binaire mais très efficace en régression
- [G] *Aucune de ces réponses n'est correcte.*

Question 7 ♣ On considère $\widehat{m}$ un estimateur qui souffre de surapprentissage. Cocher la (ou les) assertion(s) vraie(s) :

- ■ $\widehat{m}$ possède un biais faible
- B $\widehat{m}$ possède une variance faible
- C Les données d'un échantillon test seront très bien prédites par $\widehat{m}$
- ■ Les données d'apprentissage seront très bien ajustées par $\widehat{m}$
- E *Aucune de ces réponses n'est correcte.*

Question 8 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- ■ Le nombre de composantes PCR peut se choisir en minimisant l'erreur quadratique calculée par validation croisée
- ■ Lorsque p est grand, un estimateur à p composantes PCR risque de surajuster les données
- C On doit toujours prendre le plus grand nombre de composantes PLS
- D On doit toujours prendre le plus grand nombre de composantes PCR
- ■ L'estimateur PCR à p composantes est semblable à l'estimateur des MCO
- F *Aucune de ces réponses n'est correcte.*

Question 9 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- ■ Les observations y_i sont utilisées pour calculer les poids de la première composante PLS
- ■ Les observations x_i sont utilisées pour calculer les poids de la première composante PLS
- ■ L'estimateur PCR (de $m(x)$) à une composante s'écrit comme une combinaison linéaire des variables explicatives
- D Les observations y_i sont utilisées pour calculer les poids de la première composante PCR
- ■ L'estimateur PLS (de $m(x)$) à deux composantes s'écrit comme une combinaison linéaire des variables explicatives
- ■ Les observations x_i sont utilisées pour calculer les poids de la première composante PCR
- G L'estimateur PLS à une composante correspond à l'estimateur PCR à p composantes
- H *Aucune de ces réponses n'est correcte.*

Question 10 Lors d'une régression PCR, la première composante principale est la composante dont le produit scalaire avec $\mathbb{Y} = (y_1, \ldots, y_n)$ est

A maximum B minimum ■ aucune réponse n'est correcte

Question 11 ♣ On considère les estimateurs ridge définis par la pénalité proposée à la question 14. Cocher la (ou les) assertion(s) vraie(s) :

- ■ L'AUC peut être utilisé pour choisir λ dans un problème de classification binaire
- B L'erreur de classification peut être utilisée pour choisir λ dans un problème de régression
- C L'AUC peut être utilisé pour choisir λ dans un problème de régression
- ■ L'erreur quadratique de prévision peut être utilisée pour choisir λ dans un problème de régression
- E *Aucune de ces réponses n'est correcte.*

Question 12 ♣ On considère les estimateurs ridge définis par la pénalité proposée à la question 14. Cocher la (ou les) assertion(s) vraie(s) :

- [A] Il faut toujours choisir λ le plus grand possible
- [B] Il faut toujours choisir λ le plus petit possible
- [■] Les estimateurs obtenus seront proches de 0 pour de très grandes valeurs de λ
- [D] Les estimateurs obtenus seront proches des estimateurs MCO pour de très grandes de λ
- [■] Les estimateurs obtenus seront proches des estimateurs MCO pour de très petites valeurs de λ
- [F] Les estimateurs obtenus seront proches de 0 pour de très petites valeurs de λ
- [G] *Aucune de ces réponses n'est correcte.*

Question 13 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- [■] Les méthodes régularisées de type lasso/ridge permettent de réduire la variance des estimateurs MCO
- [■] Le biais (au carré) des estimateurs ridge est supérieur ou égal à celui des MCO
- [C] Le biais (au carré) des estimateurs ridge est plus petit que celui des MCO
- [D] Les estimateurs ridge/lasso sont toujours plus performants que les estimateurs des MCO
- [E] Les méthodes régularisées de type lasso/ridge permettent de réduire le biais des estimateurs MCO
- [■] On utilise généralement les méthodes ridge/lasso lorsque p est grand
- [G] *Aucune de ces réponses n'est correcte.*

Question 14 Soit $\lambda \geq 0$ et $\beta_0, \beta_1, \ldots, \beta_p$ les paramètres du modèle linéaire. Les estimateurs ridge s'obtiennent en pénalisant le critère des moindres carrés par

- [A] $\lambda \sum_{j=1}^{p} |\beta_j|$
- [B] $\lambda \sum_{j=1}^{p} \log(|\beta_j|)$
- [C] $\lambda \sum_{j=1}^{p} \sqrt{\beta_j}$
- [■] $\lambda \sum_{j=1}^{p} \beta_j^2$
- [E] $\lambda \sum_{j=1}^{p} \log(\beta_j^2)$
- [F] aucune réponse n'est correcte

Question 15 Soit $\lambda \geq 0$ et $\beta_0, \beta_1, \ldots, \beta_p$ les paramètres du modèle linéaire. Les estimateurs lasso s'obtiennent en pénalisant le critère des moindres carrés par

- [A] $\lambda \sum_{j=1}^{p} \log(|\beta_j|)$
- [B] $\lambda \sum_{j=1}^{p} \log(\beta_j^2)$
- [C] $\lambda \sum_{j=1}^{p} \beta_j$
- [D] $\lambda \sum_{j=1}^{p} \sqrt{\beta_j}$
- [E] $\lambda \sum_{j=1}^{p} \beta_j^2$
- [■] aucune réponse n'est correcte

Question 16 ♣ Cocher la (ou les) assertion(s) vraie(s) :

- [A] La norme 1 des estimateurs lasso est toujours plus grande que celle des MCO (lorsqu'on utilise les mêmes données)
- [B] Ridge permet de faire de la sélection de variables contrairement au lasso
- [■] La norme 2 des estimateurs ridge est toujours plus petite que celle des MCO (lorsqu'on utilise les mêmes données)
- [■] Ridge et Lasso permettent de réduire la complexité du modèle de régression et de répondre au problème du surapprentissage des MCO
- [E] Ridge permet d'augmenter la complexité du modèle de régression en cas de sous-apprentissage
- [F] Le lasso permet d'augmenter la complexité du modèle de régression en cas de sous-apprentissage
- [G] *Aucune de ces réponses n'est correcte.*

Question 17 ♣

- ■ La fonction **glmnet** permet de calculer les estimateurs des régressions ridge.
- ■ La fonction **glmnet** permet de calculer les estimateurs des régressions lasso.
- C La fonction **glmnet** permet de sélectionner le paramètre de régularisation des régressions lasso.
- D La fonction **glmnet** permet de faire des régressions PCR.
- ■ L'argument alpha de la fonction **glmnet** permet d'indiquer si on souhaite faire du ridge ou du lasso.
- F La fonction **glmnet** permet de sélectionner le paramètre de régularisation des régressions ridge.
- G L'argument lambda des fonctions **glmnet** et **cv.glmnet** permet d'indiquer si on souhaite faire du ridge ou du lasso.
- ■ La fonction **cv.glmnet** permet de sélectionner le paramètre de régularisation des régressions ridge.
- ■ La fonction **cv.glmnet** permet de sélectionner le paramètre de régularisation des régressions lasso.
- J *Aucune de ces réponses n'est correcte.*

Question 18 ♣ On effectue une validation croisée pour sélectionner le paramètre λ de l'algorithme lasso avec **cv.glmnet**. Cette méthode permet d'obtenir 2 valeurs que l'on identifie dans **R** par lambda.min et lambda.1se. Cocher la (ou les) assertion(s) vraie(s) :

- ■ Utiliser lambda.1se permet d'obtenir un modèle plus parcimonieux qu'avec lambda.min
- B Le nombre de variables dans le modèle utilisant lambda.min est toujours plus petit ou égal à au nombre de variables dans le modèle qui utilise lambda.1se
- ■ lambda.1se $\geq$ lambda.min
- D L'erreur calculée par validation croisée est plus petite avec lambda.1se qu'avec lambda.min
- E lambda.1se $\leq$ lambda.min
- F *Aucune de ces réponses n'est correcte.*

Question 19 L'argument α (alpha) de la fonction **glmnet** correspond à la pénalité (on ne fait pas figurer la constante de régularisation λ dans les réponses, on s'intéresse uniquement à la partie qui concerne α) :

- A $(1-\alpha)\sum_{j=1}^{p}\beta_j + \alpha\sum_{j=1}^{p}\beta_j^2$
- B $\alpha\sum_{j=1}^{p}\beta_j + (1-\alpha)\sum_{j=1}^{p}\beta_j^2$
- C $\alpha\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2$
- D $\frac{\alpha}{1-\alpha}\sum_{j=1}^{p}|\beta_j| + \frac{(1-\alpha)}{\alpha}\sum_{j=1}^{p}\beta_j^2$
- E $\frac{(1-\alpha)}{2}\sum_{j=1}^{p}|\beta_j| + 2\alpha\sum_{j=1}^{p}\beta_j^2$
- F $2\alpha\sum_{j=1}^{p}|\beta_j| + \frac{(1-\alpha)}{2}\sum_{j=1}^{p}\beta_j^2$
- G $(1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2$
- ■ $\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2$

Feuille de réponses :

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

⟵ codez votre numéro d'étudiant ci-contre, et inscrivez votre nom et prénom ci-dessous.

Nom et prénom :

.....

.....

Les réponses aux questions sont à donner exclusivement sur cette feuille : les réponses données sur les feuilles précédentes ne seront pas prises en compte.

Question 1 : ■ B C

Question 2 : A ■ C

Question 3 : A ■ C ■ E

Question 4 : ■ ■ C ■ E F ■ H I

Question 5 : ■ B C D E F G H ■ J K

Question 6 : A B ■ ■ E F G

Question 7 : ■ B C ■ E

Question 8 : ■ ■ C D ■ F

Question 9 : ■ ■ ■ D ■ ■ G H

Question 10 : A B ■

Question 11 : ■ B C ■ E

Question 12 : A B ■ D ■ F G

Question 13 : ■ ■ C D E ■ G

Question 14 : A B C ■ E F

Question 15 : A B C D E ■

Question 16 : A B ■ ■ E F G

Question 17 : ■ ■ C D ■ F G ■ ■ J

Question 18 : ■ B ■ D E F

Question 19 : A B C D E F G ■