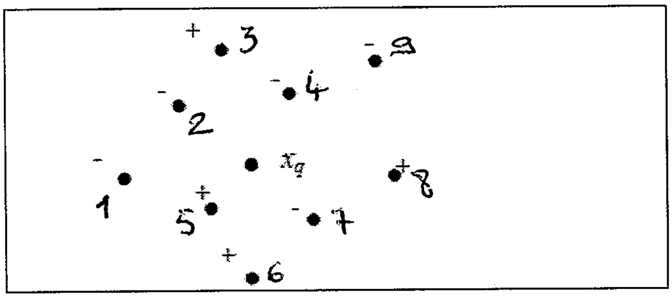
Université d'Angers
Département de Mathématiques

Année 2022-2023
M2 Data Science

Apprentissage Statistique en Grande Dimension- Examen

Le barème est approximatif. On veillera à justifier les réponses avec soin.

Exercice 1 (4,5 pts). On considère le jeu de données dans $\mathbb{R}^2$ représenté dans la figure ci-dessous. On dispose de 9 points d'entrainement et d'un point x_q à prédire. Dans cet exercice, la fonction de perte



considérée est l'erreur de classification.

1. (0,5 pt) Déterminez la prédiction associée au point x_q dans le cadre d'un 1-plus proche voisin et d'un 3-plus proche voisin.
2. (1 pt) Calculez l'erreur de validation croisée 2-folds lorsque les points 1/2/3/4, 5/6/7/8 forment les 2 groupes et que l'on considère l'algorithme des 3-plus proches voisins.
3. (1 pt) Quel est le nombre minimal de vecteurs supports que peut contenir un SVM linéaire construit sur les 9 points d'apprentissage ? Expliquez votre réponse.
4. (1 pt) En indiquant les abscisses et ordonnées par des x_i et y_j, fabriquer un arbre binaire générant une erreur d'entraînement nulle.
5. (1 pt) Calculez l'hétérogénéité globale d'un arbre construit à partir d'une seule coupe laissant les points 2/3/4/9 ensemble, lorsque la "métrique" est l'indice de Gini.

Exercice 2 (Classification binaire, 9 pts). On suppose que (X, Y) est un couple de variables aléatoires tel que $X \in \mathcal{X}$ et $Y \in \{-1, 1\}$. On se donne une fonction $\phi : \mathbb{R} \to \mathbb{R}$ et on définit pour $f : \mathcal{X} \to \mathbb{R}$ le risque R_f par :

$$R_f = \mathbb{E}[\phi(Yf(X))].$$

On note $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

1. (1 pt) Exprimez R_f sous la forme $\mathbb{E}[\Psi(X)]$ où Ψ est une fonction à déterminer.
2. (2 pts) On suppose que $\phi(-1) > \phi(1)$. Montrez avec soin que
$$\min_{f:\mathbb{R}\to\{-1,1\}} R_f = \mathbb{E}[\min\{\eta(X)(\phi(1) - \phi(-1)) + \phi(-1), \eta(X)(\phi(-1) - \phi(1)) + \phi(1)\}],$$
et que ce minimum est atteint en la règle de Bayes usuelle (associée à l'erreur de classification).
3. (1,5 pt) On suppose que X suit la loi uniforme sur $[0, 1]$, que $\phi(x) = \exp(-x)$ et que $\eta(x) = x$. Montrez que dans ce cas, le risque optimal satisfait :
$$\min_{f:\mathbb{R}\to\{-1,1\}} R_f = -\frac{1}{2}\mathrm{sh}(1) + \mathrm{ch}(1).$$

4. A partir de maintenant, on suppose que l'on dispose d'une famille de fonctions $(f_\theta)_\theta$ de $\mathcal{X}$ dans $\mathbb{R}$ et que l'on cherche
$$\theta^\star = \mathrm{Argmin}_{\theta\in\Theta} R_{f_\theta}$$
où $\theta^\star$ est supposé exister et être unique pour simplifier.
   (a) (1 pt) Supposons que l'on dispose d'un échantillon $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ issu de (X, Y). Définir un estimateur $\hat{\theta}_N$ de $\theta^\star$ (On pensera à considérer le minimiseur d'une fonction de l'échantillon naturellement associée à R_{f_θ}).
   (b) (1,5 pt + 0,5 pt bonus) On suppose dans cette question que $\phi(x) = e^{-x}$ et on définit l'algorithme suivant :
$$\theta_{k+1} = \theta_k + \frac{\gamma}{N}\sum_{i=1}^{N} \nabla_\theta f_{\theta_k}(X_i) Y_i e^{-Y_i f_{\theta_k}(X_i)}.$$
   Expliquez à quoi correspond cet algorithme et pourquoi il a vocation à converger vers $\hat{\theta}_N$ lorsque $k \to +\infty$. Cette convergence a-t-elle lieu lorsque $f_\theta(x) = \langle x, \theta\rangle$?
   (c) (1 pt) Ecrivez la version mini-batch (stochastique) de l'algorithme ci-dessus.
   (d) (1 pt) Notons Θ, l'espace des paramètres associé aux fonctions f_θ. Quels types de problèmes peut-on rencontrer en fonction de la taille Θ ? (en quelques lignes).

Exercice 3 (6,5 points). On note $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ un échantillon d'apprentissage (*i.i.d.*) de taille N tel que pour tout $1 \leq i \leq N$, $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. On note $\mathcal{L}$ la fonction définie par
$$\mathcal{L}(\theta) = \sum_{i=1}^{N} (Y_i - \langle X_i, \theta\rangle)^2 + \lambda \sum_{j=1}^{p-1} |\theta_{j+1} - \theta_j|,$$
où λ est un réel positif (Il s'agit d'un Fused-Lasso simplifié). On note (lorsque celui-ci est bien défini)
$$\hat{\theta} = \mathrm{Argmin}_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta).$$

1. (0,5 pt) Quel sens donnez-vous à $\hat{\theta}$? Quel type d'application cela peut-il avoir ?
2. (1 pt) Montrez que le sous-gradient associé à $\mathcal{L}$ existe puis calculez $\partial_{\theta_j}\mathcal{L}$ pour $j = 2, \ldots, p-1$ (On pensera à séparer les cas).
3. (1,5 pt) On suppose que $\hat{\theta}_1 < \hat{\theta}_2 < \ldots < \hat{\theta}_p$ et que $\mathbf{X}^T\mathbf{X} = I_p$. Déterminez l'expression de $\hat{\theta}_j$ pour $j = 2, \ldots, p-1$ dans ce cas.
4. (1,5 pt) On suppose dans cette question que $|\mathbf{X}_j^T\mathbf{Y}| \leq \frac{\lambda}{N}$ pour $j = 2, \ldots, p-1$, que $|\mathbf{X}_1^T\mathbf{Y}| \leq \frac{\lambda}{2N}$ et que $|\mathbf{X}_p^T\mathbf{Y}| \leq \frac{\lambda}{2N}$. Montrez qu'alors $0_{\mathbb{R}^p}$ est un minimum de $\mathcal{L}$ (On rappelle que $\mathbf{X}$ est la matrice ayant pour lignes $X_1, \ldots, X_N$ et colonnes $\mathbf{X}_1, \ldots, \mathbf{X}_p$, et $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$).
5. Dans cette question, on modifie légèrement la fonction $\mathcal{L}$ en posant
$$\tilde{\mathcal{L}}(\theta) = \sum_{k=1}^{N} (Y_i - \langle X_i, \theta\rangle)^2 + \lambda \left(\sum_{j=1}^{p-1} |\theta_{j+1} - \theta_j| + |\theta_p| \right).$$
   (a) (0,5 pt) Montrez que $\sum_{j=1}^{p-1} |\theta_{j+1} - \theta_j| + |\theta_p| = \|D\theta\|_1$ où D est une matrice que l'on déterminera.
   (b) (1,5 pt) Supposons que $(\mathbf{X}D^{-1})^T\mathbf{X}D = I_p$. En s'appuyant sur le cours et sur un changement de variable, déterminez l'expression explicite de $\hat{\theta}$ dans ce cas (où ici $\hat{\theta} = \mathrm{Argmin}_{\theta\in\mathbb{R}^p}\tilde{\mathcal{L}}(\theta)$).

Correction Examen

Exercice 1:

1. . $\hat{f}_{1ppv}(x_q) = +$ (car le plus proche voisin est affecté d'un ⊕).

   . $\hat{f}_{3ppv}(x_q) = -$ (car 2 des 3ppv sont ⊖).

2. Sur 1/2/3/4, on fabrique un prédicteur $\hat{f}_1$.
   Sur 5/6/7/8, on fabrique un prédicteur $\hat{f}_2$.

Comme c'est un 3ppv, on a alors : . $\hat{f}_1(x) = ⊖ \quad \forall x$

. $\hat{f}_2(x) = ⊕ \quad \forall x$.

De sorte que $\hat{f}_1$ se trompe 3 fois sur 4 sur 5/6/7/8.
$\hat{f}_2$ ——————— sur 1/2/3/4.

$$\Rightarrow \hat{R}_{CV} = \frac{1}{2}\left(\frac{3}{4} + 3/4\right) = 3/4.$$

3. Points non linéairement séparables, au moins 2 points du mauvais côté de l'hyperplan ⟹ 4 vecteurs supports au minimum, les 2 mal placés et les 2 que l'on trouve déjà dans une séparation de points linéairement séparable.

4. Non traité.

5. On rappelle que l'hétérogénéité globale est ici :

$$H = 2\left[N_A(\hat{p}_A(1-\hat{p}_A)) + \underbrace{(N - N_A)}_{N_B}(\hat{p}_B(1-\hat{p}_B))\right)$$

avec A, B constituant les 2 feuilles, N_A le nb d'individus dans la feuille A, $\hat{p}_A$ la proportion de + dans A, $\hat{p}_B$ la proportion de – dans B. Si on fait une coupe horizontale laissant les points 2/4/3/9 dans A, on obtient : $\frac{H}{2} = 4 \times \frac{1}{4} \times \frac{3}{4} + 5 \times \frac{3}{5} \times \frac{2}{5} = \frac{3}{4} + \frac{6}{5} = \boxed{\frac{9}{5}}$

Exercice 2:

1. $R_f = \mathbb{E}[\phi(Y.f(X))] = \mathbb{E}[\mathbb{E}[\phi(Y.f(X)) \mid X]]$

$= \mathbb{E}[\phi(f(X)).P(Y=1 \mid X)] + \mathbb{E}[\phi(f(-X)).P(Y=-1 \mid X)]$

$= \mathbb{E}[\phi(f(X))\eta(X) + \phi(f(-X))(1-\eta(X))]$

$= \mathbb{E}[\Psi(X)]$ avec $\Psi(x) = \phi(f(x))\eta(x) + \phi(f(-x))(1-\eta(x))$.

2. Comme f est à valeurs dans $\{-1, 1\}$

$\Psi(x) \in \{\phi(1)\eta(x) + (1-\eta(x))\phi(-1), \phi(-1)\eta(x) + (1-\eta(x))\phi(1)$

$\eta(x)(\phi(1)-\phi(-1)) + \phi(-1), \eta(x)(\phi(-1)-\phi(1)) + \phi(1)\}$

$\Rightarrow \Psi(x) \geq \min\{\eta(x)(\phi(1)-\phi(-1)) + \phi(-1), \eta(x)(\phi(-1)-\phi(1)) + \phi(1)\}$

$\Rightarrow R_f \geq \mathbb{E}[\min\{\eta(X)(\phi(1)-\phi(-1)) + \phi(-1), \eta(X)(\phi(-1)-\phi(1)) + \phi(1)\}$

Il nous reste à montrer que ce minorant est un minimum.

Posons $a = -\phi(1) + \phi(-1)$. On constate que $\forall \eta \in [0,1]$,

(*) $-a\eta + \phi(-1) \geq a\eta + \phi(1) \iff 2a\eta \leq a \iff \eta \leq 1/2$

car $a > 0$ par hypothèse. Ainsi, si l'on pose

$$f^*(x) = \begin{cases} 1 & \text{si } \eta(x) \geq 1/2 \\ -1 & \text{si } \eta(x) < 1/2 \end{cases},$$

on constate que
$$\Psi(x) = \begin{cases} \phi(1)\eta(x) + \phi(-1)(1-\eta(x)) & \text{si } \eta(x) \geq 1/2 \\ \phi(-1)\eta(x) + \phi(1)(1-\eta(x)) & \text{si } \eta(x) \leq 1/2 \end{cases}$$

$$\overset{(**)}{=} \begin{cases} -a\eta(x) + \phi(-1) & \text{si } \eta(x) \geq 1/2 \\ a\eta(x) + \phi(1) & \text{si } \eta(x) \leq 1/2 \end{cases}$$

$$= \min\{-a\eta(x) + \phi(-1), a\eta(x) + \phi(1)\}$$

d'après (*). Il vient $R_{f^*} = \mathbb{E}[\min\{-a\eta(X) + \phi(-1), a\eta(X) + \phi(1)\}]$

puis le résultat.

3. Si $\eta(x) = x$, alors en posant $a = e - e^{-1}$

$$\Psi_{f^*}(x) = \begin{cases} ax + e^{-1} & \text{si } x \leq 1/2 \\ -ax + e & \text{si } x \geq 1/2 \end{cases} \quad \text{(d'après (**))}$$

Ainsi, $R_{f^*} = \mathbb{E}[\Psi_{f^*}(X)]$

$$= a\,\mathbb{E}[X \mathbb{1}_{\{X \leq 1/2\}}] + e^{-1} P(X \leq 1/2)$$
$$- a\,\mathbb{E}[X \mathbb{1}_{\{X \geq 1/2\}}] + e\, P(X \geq \tfrac{1}{2})$$

$\mathbb{E}[X \mathbb{1}_{\{X \leq 1/2\}}] = \int_0^{1/2} x\,dx = \left[\frac{x^2}{2}\right]_0^{1/2} = \frac{1}{8}$ et $\mathbb{E}[X\mathbb{1}_{\{X \geq 1/2\}}] = \left[\frac{x^2}{2}\right]_{1/2}^{1} = \frac{3}{8}$

de sorte que :

$$R_{f^*} = -\frac{1}{4}a + \frac{e + e^{-1}}{2} = \boxed{-\frac{1}{2}\,\mathrm{sh}(1) + \mathrm{ch}(1).}$$

4. (a) Il est naturel de poser :

$$\hat{\theta}_N = \operatorname{Argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{k=1}^{N} \phi(Y_k f_\theta(X_k)).$$

car par la LGN, $\frac{1}{N} \sum_{k=1}^{N} \phi(Y_k \cdot f_\theta(X_k)) \xrightarrow{N \to +\infty} \mathbb{E}[\phi(Y \cdot f_\theta(X))]$

[si $\mathbb{E}[|\phi(Y \cdot f_\theta(X))|] < +\infty$)

(b) Comment calculer $\hat{\theta}_n$? L'idée proposée ici est une simple descente de gradient de pas γ. En effet, comme

$$\nabla_\theta e^{-y \cdot f_\theta(x)} = -y \nabla_\theta f_\theta(x) e^{-y \cdot f_\theta(x)}, \text{ on constate}$$

que (θ_k) est bien de la forme :

$$\theta_{k+1} = \theta_k - \gamma \times \nabla_\theta \left(\frac{1}{N} \sum_{k=1}^{N} e^{-Y_k f_{\theta_k}(X_i)} \right)$$

. A propos de la convergence lorsque $f_\theta(x) = \langle x, \theta \rangle$.

. La descente de gradient converge vers θ^* lorsque la fonction à minimiser est convexe avec pour unique minimum θ^*. Ici, le fait que θ^* soit l'unique minimum est admis (Pas toujours vrai évidemment). Il nous reste pour vérifier la convergence vers θ^* à montrer que $\theta \to \frac{1}{N} \sum_{k=1}^{N} e^{-Y_i f_\theta(X_k)}$ est convexe.

Or, $x \to \langle x, \theta \rangle$ est convexe (car linéaire) et $u \to e^{\lambda u}$ est convexe $\forall \lambda \in \mathbb{R}$ $\Rightarrow$ $\theta \to e^{-y f_\theta(x)}$ est convexe $\forall (x,y)$.

Le résultat suit.

(c) Pour une version mini-batch, il suffit de faire un tirage aléatoire $(I_1^{(k)}, \dots, I_M^{(k)})$ de M observations $(M \ll N)$ à chaque itération et d'écrire :

$$\theta_{k+1}^{(MB)} = \theta_k^{(MB)} + \frac{\gamma}{M} \sum_{i=1}^{M} \nabla_\theta f_{\theta_k}^{(MB)}(X_{I_i^{(k)}}) Y_{I_i^{(k)}} e^{-Y_{I_i^{(k)}} f_{\theta_k}(X_{I_i^{(k)}})}$$

(d) Classiquement, si Θ est "grand" (de grande dimension) la convergence de l'algorithme de descente sera plus lente mais surtout, on se confronte à un risque de surapprentissage(*) lié à une variance trop élevée de $\hat{\theta}_N$. Si Θ est "petit", la variance est plus faible mais le biais est plus important.

(*) : Dans le cadre des réseaux de neurones, on constate néanmoins qu'il n'y a pas autant de surapprentissage que ce que la dimension de Θ devrait générer. Il est "communément" admis que l'algorithme choisit des minimiseurs qui sont de "bonnes solutions"...

Exercice 3

1) Il s'agit d'un problème de régression pénalisé avec une pénalisation L^1 sur les accroissements. On veut donc faire en sorte que peu d'accroissements soient $\neq$ de 0 (La pénalisation L^1 a vocation à approximer la pénalisation L^0 qu'elle même est conçue pour générer de la parcimonie).

Applications : Données "géographiques" (Météo, Capteurs sur un cerveau...)

2) $\theta \to |\theta_{j+1} - \theta_j|$ est clairement convexe (par exemple, $\forall \lambda \in [0,1]$ on a bien : $|(\lambda\theta+(1-\lambda)\tilde{\theta})_{j+1} - (\lambda\theta+(1-\lambda)\tilde{\theta})_j|$
$$\leq \lambda|\theta_{j+1}-\theta_j| + (1-\lambda)|\tilde{\theta}_{j+1}-\tilde{\theta}_j|).$$

Ainsi, $\mathcal{L}$ est convexe en tant que somme de fonctions convexes.

Notons $\mathbb{X} = (X_1|\dots|X_j)$. $\forall j \in \{2,\dots,p-1\}$,

$$\partial_{\theta_j}\mathcal{L} = 2N\,\mathbb{X}_j^T(\mathbb{X}\theta - Y) + \lambda\begin{cases} \operatorname{sgn}(\theta_{j+1}-\theta_j)+\operatorname{sgn}(\theta_j-\theta_{j-1}) & \text{si } \theta_j\neq\theta_{j+1},\ \theta_{j-1}\neq\theta_j \\ \operatorname{sgn}(\theta_{j+1}-\theta_j)+[-1,1] & \text{si } \theta_j\neq\theta_{j+1},\ \theta_{j-1}=\theta_j \\ [-1,1]+\operatorname{sgn}(\theta_j-\theta_{j-1}) & \text{si } \theta_j=\theta_{j+1},\ \theta_{j-1}\neq\theta_j \\ 2[-1,1] & \text{si } \theta_{j-1}=\theta_j=\theta_{j+1} \end{cases}$$

3) Si $\mathbb{X}^T\mathbb{X} = I_p$, alors $\mathbb{X}_j^T\mathbb{X}\theta = \theta_j$.

Si $\hat{\theta}_1 < \dots < \hat{\theta}_p$, alors, il vient : $\forall j \in [\![2, p-1]\!]$
$$\partial_{\theta_j}\mathcal{L}(\hat{\theta}) = 0 = 2N(\hat{\theta}_j - X_j^TY) + 2\lambda.$$
$$\Rightarrow \forall j \in [\![2, p-1]\!], \quad \hat{\theta}_j = X_j^TY - \frac{\lambda}{N}$$

4) Sous ces conditions on constate que
$$-2N\,X_j^TY + 2\lambda[-1,1] \quad \text{contient bien } 0 \ \forall j \in [\![2,p-1]\!]$$
Pour les cas $j=1$, $j=p$, il y a une petite différence dans le

sous-gradient et il faut alors que :

$$-2N\, X_j^T Y + \lambda[-1,1] \text{ contienne } 0 \quad (j=1 \text{ ou } p).$$

C'est bien le cas à nouveau sous les conditions de la question.

Il vient $0 \in \{\theta,\ \underset{\text{sous-gradient}}{\nabla \mathcal{L}(\theta)} = 0\} \Rightarrow 0$ est minimum

via un résultat du cours.

5. (a)

$$\begin{pmatrix} \theta_2-\theta_1 \\ \theta_3-\theta_2 \\ \vdots \\ \theta_p-\theta_{p-1} \\ \theta_p \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & -1 & 1 \\ 0 & & & 1 \end{pmatrix}}_{D} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$$

$\Rightarrow$ résultat.

(b) On a donc à chercher "le" minimum de

$$\tilde{\mathcal{L}}(\theta) = \|Y - X\theta\|^2 + \lambda \|D\theta\|_1.$$

$$\underset{=}{\beta = D\theta} \quad \|Y - XD^{-1}\beta\|^2 + \lambda\|\beta\|_1$$

(D est clairement inversible).

On a donc un problème de type Lasso classique avec $\tilde{X} = XD^{-1}$ à la place de X. Sous les hypothèses, $\tilde{X}^T X = I_p$. On sait alors que la solution (voir cours dans une forme légèrement différente en terme de normalisation) est : $\hat{\beta}^{lasso} = (\hat{\beta}_j^{lasso})_{j=1}^p$ avec

$$\hat{\beta}_j^{lasso} = \text{sgn}(\hat{\beta}_j)\left(|\hat{\beta}_j| - \frac{\lambda}{2N}\right)_+$$

où $\hat{\beta} = \tilde{X}^T Y$. On a alors $\hat{\theta} = D^{-1} \cdot \hat{\beta}$.