Université d'Angers
Département de Mathématiques

Année 2020-2021
M2 Data Science

Apprentissage Statistique en Grande Dimension- Examen

Le barême est approximatif. Le dernier exercice est dédié à la partie du cours de M. Graczyk. Il devra être effectué sur une copie séparée.

Exercice 1 (6pts)**.** Soit (X, Y) une variable aléatoire à valeurs dans $\mathcal{X} \times \{0, 1\}$. On note $\eta(x) := \mathbb{P}(Y = 1|X = x)$.

1. (2pts) On suppose que η est défini par

$$\eta(x) = \frac{x}{x+\theta}$$

où $\theta \in]0, 1[$ et que X suit la loi uniforme sur l'intervalle $[0, \alpha\theta]$ où $1 < \alpha < 1/\theta$. Montrez que le risque de Bayes associé à l'erreur de classification standard ($\ell(y, y') = 1_{\{y \neq y'\}}$) est égal à

$$\frac{1}{\alpha}\left(1 + \ln\left(\frac{1+\alpha}{4}\right)\right).$$

2. On suppose dans cette question que $\ell(y, y') = 1_{\{y \neq y'\}} + \lambda 1_{\{y'=1\}}$ avec $\lambda > 0$.
   (a) (2pts) Soit g une fonction de $\mathcal{X}$ vers $\{0, 1\}$. Montrez que

$$L(g) = \mathbb{E}[\ell(Y, g(X))] = \mathbb{E}\left[(1 + \lambda - \eta(X))1_{\{g(X)=1\}} + \eta(X)1_{\{g(X)=0\}}\right].$$

   Indication : on pourra s'inspirer de la preuve dans le cas classique.

   (b) (2pts) En déduire $g^\star = \text{Argmin}_{g:\mathcal{X}\to\{0,1\}} L(g)$ et montrer que le risque optimal est égal à

$$\mathbb{E}[\min(1 + \lambda - \eta(X), \eta(X)].$$

Exercice 2 (7pts)**.** 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

$$\mathcal{L}(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \Psi(Y_i, \langle X_i, \theta\rangle) + \lambda(\alpha\|\theta\|_1 + (1-\alpha)\frac{\|\theta\|_2^2}{2}).$$

On note

$$\hat{\theta} = \text{Argmin}_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta).$$

1. (1,5 pt) On suppose dans cette question que $Y_i \in \{0,1\}$ et que

$$\Psi(y,z) = y\frac{e^z}{1+e^z} + (1-y)\frac{1}{1+e^z}.$$

Expliquez à quel type d'estimateur correspond $\hat{\theta}$ et quelle fonction `R` vous pouvez utiliser pour calculer $\hat{\theta}$? (On précisera les paramètres de la fonction `R` à utiliser. Dans le cas où vous ne souviendriez pas de la formule exacte, vous pourrez simplement indiquer quels types d'informations doivent être fournies dans cette fonction).

2. On suppose dans cette question que

$$\Psi(y,z) = (y - \phi(z))^2$$

où $\phi : \mathbb{R} \mapsto \mathbb{R}$ est une fonction convexe $\mathcal{C}^1$.

(a) (0,5pt) Sans calculs, montrez que $\mathcal{L}$ est convexe.

(b) (1,5pt) Pour tout j, calculez $\partial_{\theta_j}\mathcal{L}$ (au sens sous-gradient, on distinguera donc le cas $\theta_j \neq 0$ du cas $\theta_j = 0$).

(c) (1,5pt) On suppose maintenant que $\phi(x) = x$. Ecrire l'équation de "stationnarité", *i.e.* donnez une condition nécessaire et suffisante sur θ pour que le minimum soit atteint en ce point.

(d) (2pts) On suppose toujours que $\phi(x) = x$. Notons $\mathbf{x}$ la matrice ayant pour lignes $X_1,\ldots,\ X_N$. Déduisez de la question précédente que si $\mathbf{x}^T\mathbf{x} = I_p$, alors pour tout $j \in \{1,\ldots,p\}$,

$$\hat{\theta} = \frac{\mathrm{sgn}(\mathbf{x}_j^T Y)}{1+N\lambda(1-\alpha)}(|\mathbf{x}_j^T Y| - N\lambda\alpha)_+,$$

où $\mathbf{x}_j$ désigne la j-ième colonne de $\mathbf{x}$.

Exercice 3 (2,5pts+Bonus)**.** On note $\{(X_1,Y_1),\ldots(X_n,Y_n)\}$ un échantillon d'apprentissage à valeurs dans $\mathcal{X}\times\mathbb{R}$ (Y est quantitative). Soient B échantillons bootstrap de taille $m \leq n$ et on suppose que l'on fabrique ainsi B prédicteurs notés $\hat{f}_1,\ldots,\ \hat{f}_B$ avec la même règle de construction (par exemple un k-ppv).

1. (1pt) Soit $x \in \mathcal{X}$. Expliquez avec soin pourquoi la loi de $\hat{f}_b(x)$ et la covariance entre $\hat{f}_{b_1}(x)$ et $\hat{f}_{b_2}(x)$ ne dépend pas des indices choisis.
2. (1,5pt) Calculez la variance du prédicteur "baggé" issu des B prédicteurs précédents en fonction de paramètres que l'on précisera.
3. (Bonus) Pensez-vous que dans de mauvaises situations, le bagging puisse altérer la qualité du prédicteur de "base" ? On pourra justifier de manière "approximative" (en introduisant les quantités adéquates).

Exercice 4 (5pts)**.** Soit un caractère statistique gaussien de dimension 3 centré $X = \big(X_1, X_2, X_3\big)^T \sim N(0, \Sigma_X)$

avec les matrices de covariance $\Sigma_X = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ et de précision $K_X = \begin{pmatrix} 0,5 & 0 & -0,5 \\ 0 & 0,5 & -0,5 \\ -0,5 & -0,5 & 2 \end{pmatrix}$.

1. Y a-t-il des composantes X_i indépendantes entre elles ?

2. Y a-t-il des composantes X_i conditionellement indépendantes sachant les autres ? Si oui, lesquelles ?
Qu'en déduit-on sur la prédiction de X_1, quand on connait X_2 et X_3 ?

3. Dessiner le graphe de dépendance $\mathcal{G}$ de X.

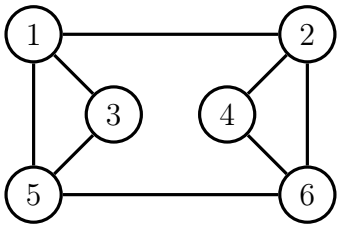
4. Déterminer la loi conditionnelle de $X_1 \,|\, (X_2 = a, X_3 = b)$.

5. Déterminer la corrélation conditionnelle $\rho_{X_2, X_3 | X_1 = u}$.

6. On sait qu'un vecteur aléatoire Y appartient au modèle graphique gaussien de dimension $p = 3$, gouverné par le graphe $\mathcal{G}$. On ne connait pas la matrice de covariance Σ_Y de Y. On a un échantillon de taille $n = 10$ de Y et on calcule la matrice de covariance empirique $\tilde{\Sigma}_Y = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. Donner l'estimateur de maximum de vraisemblance(EMV) de Σ_Y.

7. Le graphe $\mathcal{H}$ suivant est-il décomposable ? Donner sa décomposition en graphes premiers. Donner les cliques de $\mathcal{H}$ et la factorisation de la densité f du modèle gaussien gouverné par $\mathcal{H}$.

$\mathcal{H}$



Rappels. *Soit X un vecteur gaussien $N(\xi, \Sigma)$ dans $\mathbb{R}^p$ avec Σ inversible.*
On partitionne $X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}$ en sous-vecteurs $X_A \in \mathbb{R}^r$ et $X_B \in \mathbb{R}^s$, avec $r + s = p$.
La **loi conditionnelle** $X_A | \ (X_B = x_B) \ \sim N(\xi_{A|B}, \Sigma_{A|B})$ *où* $\xi_{A|B} = \xi_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \xi_B)$ *et* $\Sigma_{A|B} = K_{AA}^{-1}$.

La **corrélation conditionnelle** $\rho_{lm|V\setminus\{l,m\}} = -\tilde{\kappa}_{lm} = -\dfrac{\kappa_{lm}}{\sqrt{\kappa_{ll}}\sqrt{\kappa_{mm}}}$.

Équation de Maximum de Vraisemblance *:* $\pi_{\mathcal{G}}(\hat{K}^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma})$,
où $\tilde{\Sigma}$ est la covariance empirique.

La **décomposition d'un graphe** *est une partition (A, B, S) de l'ensemble des sommets du graphe telle que S sépare A de B et S est complet.*

Un graphe est **décomposable** *s'il est complet ou si ses composantes premières sont complètes.*

(F) *La densité* $f(\underline{x})$ *du modèle graphique gaussien avec le graphe* $\mathcal{G}$ *a la* **factorisation** $f(\underline{x}) = \prod_C \psi_C(x_C)$ *où* C *sont les cliques (sous-graphes complets maximaux) de* $\mathcal{G}$*.*