Université d'Angers
Département de Mathématiques

Année 2019-2020
M2 Data Science

Apprentissage Statistique en Grande Dimension- Examen

Dans les exercices ci-dessous, on alterne questions purement mathématiques et questions d'interprétation de certains aspects de l'apprentissage statistique en grande dimension. Dans ce deuxième cas, il est attendu d'être relativement concis dans les réponses tout en mettant l'accent sur les points qui vous semblent pertinents.

Le dernier exercice (page 4) est dédié à la partie du cours de M. Graczyk. Il devra être effectué sur une copie séparée.

Exercice 1. Soit $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un échantillon d'apprentissage où $Y_1 \in \mathcal{Y}$ où $\mathcal{Y}$ est supposé inclus dans $\mathbb{R}$ et $X_1 \in \mathcal{X} \subset \mathbb{R}^p$. Pour une fonction $g : \mathcal{X} \to \mathcal{Y}$, on note

$$L_n(g) = \frac{1}{n} \sum_{k=1}^{n} \ell(g(X_i), Y_i)$$

où ℓ est une fonction de perte. Pour une classe $\mathcal{G}$ de fonctions de $\mathcal{X}$ vers $\mathcal{Y}$, on note

$$g_n^\star = \mathrm{Argmin}_{g \in \mathcal{G}} L_n(g) \quad \text{(supposé bien défini)}.$$

1. On suppose que $\mathcal{Y} = \{1, \ldots, K\}$ et que $\ell(y, y') = 1_{\{y \neq y'\}}$.

   $$L^\star = \mathrm{Inf}_{f : \mathcal{X} \to \mathcal{Y}} L(f)$$

   où $L(f) = \mathbb{E}[\ell(f(X), Y)]$.

   (a) Que représente $L^\star$? Quelle interprétation pourriez-vous en donner ?

   (b) Montrez que $L^\star$ est un minimum et déterminez sa valeur (en le démontrant).

2. On décompose la quantité $L(g_n^\star) - L^\star$ en deux parties :

   $$L_n(g_n^\star) - L^\star = \underbrace{L_n(g_n^\star) - \inf_{g \in \mathcal{G}} L(g)}_{\mathcal{E}_1} + \underbrace{\inf_{g \in \mathcal{G}} L(g) - L^\star}_{\mathcal{E}_2}.$$

   (a) Dans quel sens varie $\mathcal{E}_2$ lorsque $\mathcal{G}$ augmente ? Expliquez.

   (b) Déterminez la limite de $\mathcal{E}_1$ lorsque $n \to +\infty$ et lorsque $\mathcal{G}$ est fini ?

   (c) Lorsque $\ell(y, y') = (y - y')^2$ et que $Y = f(X)$ où f est une fonction quelconque de $\mathbb{R}^p$ vers $\mathbb{R}$, explicitez les différentes quantités en jeu et la classe de fonctions $\mathcal{G}$ dans le cadre de la régression linéaire (On supposera les propriétés d'inversibilité satisfaites).

(d) Dans le cadre des méthodes pénalisées telles que la méthode LASSO, lorsque $\mathcal{Y} = \mathbb{R}$ et $\mathcal{X} = \mathbb{R}^p$, à quoi correspond (moralement) la classe $\mathcal{G}$?

(e) Dans le cadre des arbres de décision, quelle influence a l'augmentation de la profondeur de l'arbre sur $\mathcal{E}_1$ et $\mathcal{E}_2$?

(f) Dans le cadre des réseaux de neurones, quel type de méthode utilise-t-on pour calculer $g_n^\star$?

Exercice 2. On s'intéresse ici à l'interprétation de codes `R` :

1.
```
model <- keras_model_sequential()
model %>%
layer_dense(units = 4, input_shape = 2) %>%
layer_dropout(rate=0.4)%>%
layer_activation(activation = 'relu') %>%
layer_dense(units = 3) %>%
layer_activation(activation = 'softmax')
model %>% compile(
loss = 'categorical_crossentropy', optimizer = 'sgd'
)
```
Représentez le réseau de neurones associé et précisez les paramètres en jeu.

2. `model=svm(classe .,data=Z,kernel="polynomial",gamma=0.5,degree=2,cost=1)`
Quel type d'algorithme est-il utilisé ici ? Précisez les paramètres.

Exercice 3. On considère la fonction $L : \mathbb{R}^2 \to \mathbb{R}$ définie pour tout $\theta = (\theta_1, \theta_2)$ par

$$L(\theta) = \frac{1}{2N} \sum_{k=1}^{n} (y_k - \theta_1 x_k - \theta_2)^2 + \frac{\lambda}{2}(\|\theta\|_1 + \|\theta\|_2^2).$$

1. Pourquoi le sous-gradient de L est-il bien défini ?
2. Le calculer.
3. En déduire une condition nécessaire et suffisante qui garantisse qu'un point θ soit point critique de L.

Exercice 4 (Classification binaire (suite))**.** Soit un problème d'apprentissage où Y est à valeurs dans $\{-1, 1\}$ et $Y|X = x$ suit une loi de Rademacher de paramètre $p(x)$ où $p(x) \in [0, 1]$. On considère la fonction de perte $\ell(y, y') = \log(1 + e^{-yy'})$. Le but de l'exercice est de déterminer parmi les fonctions $f : \mathcal{X} \to \mathbb{R}\{-1, 1\}$, celle qui minimise $\mathbb{E}[\ell(Y, f(\mathbf{X}))]$.

1. Pour f fixé, écrire le risque associé à ℓ sous la forme $\mathbb{E}[\phi(p(X), f(X))]$ où ϕ est une fonction à déterminer.
2. Pour p fixé, déterminez $\min\{\phi(p, 1), \phi(p, -1)\}$ en fonction de p.
3. En déduire que le prédicteur de Bayes est la règle de décision optimale pour ce problème.

Exercice 5. Soit $\mathcal{H}$ un hyperplan affine de $\mathbb{R}^p$.

1. Rappelez pourquoi l'équation d'un hyperplan affine $\mathcal{H}$ s'écrit

$$\langle \beta, x \rangle + \beta_0 = 0.$$

2. Montrez que pour un tel hyperplan $\mathcal{H}$ la distance d'un point x à $\mathcal{H}$ est égale à :

$$d(x, \mathcal{H}) = \frac{|\langle \beta, x \rangle + \beta_0|}{\|\beta\|}.$$

Exercice 6. On considère l'entropie entre deux mesure de probabilité P et Q sur $\{1, \ldots, m\}$ définie par :

$$D(P||Q) = \sum_{k=1}^{m} P(k) \log \frac{P(k)}{Q(k)}.$$

Montrez que $D(P||Q) \geq 0$ et que $D(P||Q) = 0$ si et seulement si $P = Q$ (On pourra s'appuyer sur l'inégalité de Jensen).

Exercice 7. 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 $\Sigma_X = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ et de précision $K_X = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 1 \end{pmatrix}$.

1. Quelle est la relation entre les matrices Σ_X et K_X ?

2. Y a-t-il des composantes X_i indépendantes entre elles ? Si oui, lesquelles ?

3. 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 ?

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

5. Déterminer la loi marginale de $(X_2, X_3)^T$.

6. Déterminer la loi conditionnelle de $(X_2, X_3)^T | X_1 = u$ et la corrélation conditionnelle $\rho_{X_2, X_3 | X_1 = u}$.

7. On sait qu'un vecteur aléatoire Y appartient au modèle graphique gaussien gouverné par le graphe $\mathcal{G}$. On ne connait pas la matrice de covariance Σ_Y de Y. On a un échantillon de taille $s = 5$ de Y et on calcule la matrice de covariance empirique $\tilde{\Sigma}_Y = \begin{pmatrix} 2 & 1 & 0.9 \\ 1 & 1 & 1 \\ 0.9 & 1 & 2 \end{pmatrix}$. Donner l'estimateur de maximum de vraisemblance(EMV) de Σ_Y et l'EMV de la matrice de précision K_Y de Y.

8. Le graphe $\mathcal{G}$, est-il complet ? Décomposable ? Donner sa décomposition en cliques.

9. Donner un exemple d'un graphe non-décomposable.

Rappels. *Soit X un vecteur gaussien $N(\xi, \Sigma)$ dans $\mathbb{R}^d$ 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 $r + s = d$. On partitionne $\xi = \begin{pmatrix} \xi_A \\ \xi_B \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$, $K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}$ en bloques $\begin{pmatrix} r \times r & r \times s \\ s \times r & s \times s \end{pmatrix}$.

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.