Examen de Science des données 2023

MINES Paris - Tronc Commun 1A

Durée : 1h30. Tous documents autorisés.

QCM (5 points)

Plusieurs réponses peuvent être correctes. Aucun point n'est enlevé en cas de réponse incorrecte.

1. L'erreur quadratique moyenne d'un estimateur non-biaisé est nécessairement plus faible que celle d'un estimateur biaisé.
   A. Vrai
   B. Faux X
2. Lorsqu'on effectue une analyse en composantes principales sur un jeu de données,
   A. Il est préférable d'utiliser des variables réduites lorsque les variables ont des unités différentes X
   B. L'utilisation de variables réduites donne la même importance à toutes les variables X
3. La matrice de covariance d'un jeu de données comprenant n vecteurs de dimension d est de taille
   A. $n \times n$
   B. $n \times d$
   C. $d \times n$
   D. $d \times d$ X
4. On entraîne un modèle d'apprentissage supervisé sur un jeu de données d'entraînement $(\vec{x}^i, y_i)_{i=1}^m$. Lorsqu'on augmente m,
   A. Le biais du modèle ne change pas mais sa variance augmente
   B. Le biais et la variance du modèle diminuent
   C. Le biais et la variance du modèle restent identiques
   D. Le biais du modèle reste identique mais sa variance diminue X
5. On entraîne un modèle d'apprentissage supervisé sur un jeu de données d'entraînement $(\vec{x}^i, y_i)_{i=1}^m$. Le risque du modèle obtenu estimé sur des données de test
   A. est nécessairement plus élevé que le risque estimé sur la base d'entraînement
   B. est du même ordre de grandeur que le risque estimé sur la base d'entraînement
   C. est nul dès lors le prédicteur minimise le risque empirique sur la base d'entraînement
   D. peut être moins élevé que le risque d'entrainement. X

Exercice 1: Estimation (9 points)

Dans cet exercice, nous cherchons à estimer la loi de probabilité du nombre de véhicules s'arrêtant à un feu rouge pendant une durée T fixée. Pour ce faire, nous disposons d'un capteur placé à la hauteur du feu qui permet de mesurer le temps qui s'écoule entre le moment où le feu passe au rouge et l'arrivée de la première voiture qui s'arrête. Nous ferons l'hypothèse que le feu reste au vert suffisamment longtemps pour que toutes les voitures à l'arrêt aient le temps de circuler avant que le feu repasse au rouge.

1. Modélisation

Fixons un temps $t > 0$. On peut découper l'intervalle de temps $[0, t]$ en n segments temporels de durée Δt

$$\Delta t := \frac{t}{n}.$$

Pour tout $i = 1, \dots, n$, on note X_i la variable aléatoire correspondant au nombre de voitures qui s'arrêtent au feu au cours de l'intervalle de temps $[(i-1)\Delta t, i\Delta t]$. Pour Δt suffisamment petit, on fait l'hypothèse que:

$$X_i = \begin{cases} 1 \text{ avec probabilité } \lambda\Delta t + o(\Delta t) \\ 0 \text{ avec probabilité } 1 - \lambda\Delta t + o(\Delta t) \end{cases}$$

où $\lambda > 0$ est un paramètre inconnu.

1. *(1 point)* Montrer que la variable aléatoire

$$N_{\Delta t} = \sum_{i=1}^{n} X_i$$

suit la loi de probabilité

$$\mathbb{P}(N_{\Delta t} = k) = \binom{n}{k} (\lambda\Delta t)^k (1 - \lambda\Delta t)^{n-k} + o(1).$$

A quel évènement cette variable aléatoire correspond-elle?

> **Solution:** La variable aléatoire $N_{\Delta t}$ indique le nombre de véhicule qui arrivent au feu entre les instants 0 et t. $N_{\Delta t} = k$ si et seulement si il existe exactement k intervalles de temps $([(i_j - 1)\Delta t, i_j \Delta t])_{1 \le j \le k}$ pour lesquels $X_j = 1$. A k fixé, il existe exactement $\binom{n}{k}$ choix possibles pour ces intervalles. On en déduit donc:
>
> $$\mathbb{P}(N_{\Delta t} = k) = \binom{n}{k} (\lambda\Delta t)^k (1 - \lambda\Delta t)^{n-k} + o(1).$$

2. *(1 point)* En faisant tendre Δt vers 0 dans l'expression précédente, en déduire que le nombre $N(t)$ de véhicules s'arrêtant au feu rouge entre les instants 0 et t suit une loi de Poisson de paramètre λt:

$$\mathbb{P}(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

> **Solution:** En utilisant le fait que $\Delta t = t/n$, on vérifie que
>
> $$\mathbb{P}(N_{\Delta t} = k) = \frac{n(n-1)\dots(n-k)}{k!} \frac{(\lambda t)^k}{n^k} \left(1 - \frac{t}{n}\right)^{n-k} + o(1)$$
>
> lorsque $n \to +\infty$. Or, on vérifie que
>
> $$\left(1 - \frac{t}{n}\right)^{n-k} = \exp\left((n-k)\log\left(1 - \frac{t}{n}\right)\right) \to e^{-\lambda t}.$$
>
> lorsque $n \to +\infty$.

3. *(1 point)* On note $\mathcal{T} = \inf_{t>0}\{N(t) \geq 1\}$ le temps d'arrivée de la première voiture. Montrer que la variable aléatoire $\mathcal{T}$ suit une loi exponentielle de paramètre λ. On rappelle que la densité p d'une loi exponentielle de paramètre λ est donnée par

$$p(t) = \lambda \exp(-\lambda t).$$

Solution: On vérifie que $\mathbb{P}(\mathcal{T} > t) = \mathbb{P}(N(t) = 0)$. On en déduit que la fonction de répartition de $\mathcal{T}$ est $F(t) = 1 - \exp(-\lambda t)$. La variable aléatoire $\mathcal{T}$ suit donc bien une loi exponentielle de paramètre λ.

2. Estimation du paramètre λ

4. *(3 points)* On dispose grâce au capteur d'un échantillon $(\tau_1, \dots \tau_m)$ de réalisations de la variable aléatoire $\mathcal{T}$. Proposer un estimateur du paramètre λ à partir du principe du maximum de vraisemblance.

Solution: La vraisemblance de l'échantillon est

$$L(\tau_1, \dots \tau_n; \lambda) = \Pi_{i=1}^m \lambda \exp(-\lambda \tau_i) = \lambda^m \exp(-\lambda \sum_{i=1}^m \tau_i).$$

L'estimation du maximum de vraisemblance est donc

$$\hat{\lambda}_{\text{MSE}}^{-1} = \frac{1}{m} \sum_{i=1}^m \tau_i.$$

5. *(3 points)* On suppose maintenant que la loi de distribution *a priori* du paramètre λ est une loi gamma de paramètres α et β. On rappelle que la densité d'une loi Gamma est donnée par:

$$p(x) = \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)},$$

et que l'espérance de cette loi est α/β. Proposer un estimateur de Bayes du paramètre λ. Quelle est la valeur de cet estimateur lorsque $m = 0$? Lorsque $m \to +\infty$?

Solution: D'après la formule de Bayes, on a

$$\mathbb{P}(\lambda | \tau_1, \dots, \tau_m) \propto \mathbb{P}(\tau_1, \dots, \tau_m | \lambda) \mathbb{P}(\lambda).$$

Or, on vérifie que

$$\mathbb{P}(\tau_1, \dots, \tau_m | \lambda) \mathbb{P}(\lambda) \propto \lambda^{m+\alpha-1} \exp(-\lambda(\beta + \sum_{i=1}^m \tau_i))$$

La probabilité a posteriori est donc une loi Gamma de paramètres $\alpha' = \alpha + m$ et $\beta' = \beta + \sum_{i=1}^m \tau_i$. L'estimateur de Bayes est donc:

$$\lambda_{\text{Bayes}} = \frac{\alpha + m}{\beta + \sum_{i=1}^m \tau_i} = \frac{\alpha + m}{\beta + m \lambda_{\text{MSE}}^{-1}}.$$

Exercice 2: Boosting (6 Points)

On dispose d'un jeu de données $((\vec{x}^i, y^i) \in \mathbb{R}^p \times \{-1, 1\}, i = 1, \ldots, m)$. On cherche à entraîner un classifieur sur ce jeu de données en utilisant la méthode du *boosting*. Cette méthode vise à construire une fonction de décision f_N qui combine linéairement différents classifieurs $(h_k)_{k=1,\ldots,N}$ dits *faibles* appartenant à un espace d'hypothèse $\mathcal{F}$ fixé. Pour ce faire, on procède en minimisant le risque empirique évalué avec une fonction de coût exponentielle:

$$R(f) = \frac{1}{m} \sum_{i=1}^{m} \exp(-y_i f(\vec{x_i})).$$

On notera bien que les classifieurs faibles $(h_k)_{k=1,\ldots,N}$ de $\mathcal{F}$ sont des fonctions sur $\mathbb{R}^p$ à valeurs dans $\{-1, 1\}$, et que la fonction de décision f_N est à valeurs dans $\mathbb{R}$.

1. *(2 points)* Montrer que le risque empirique $R(f)$ calculé avec une fonction de coût exponentielle est un majorant du risque empirique calculé avec la fonction de coût 0/1.

   Solution: Le risque empirique calculé avec la fonction de coût 0/1 est

   $$R_{0/1}(f) = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \neq f(\vec{x_i})}$$

   On vérifie que $1_{y_i \neq f(\vec{x_i})} \leq \exp(-y_i f(\vec{x_i})$ pour tout i. On a donc bien $R_{O/1}(f) \leq R(f)$.

La fonction de décision finale obtenue avec la méthode du boosting est $f_N = \sum_{k=1}^{N} \alpha_k h_k$, où pour tout $k = 1, \ldots, N$, α_k est un coefficient de pondération positif et h_k un classifieur faible de $\mathcal{F}$. f_N est construit itérativement en N étapes successives. A l'initialisation, on fixe $f_1 = h_1$ où

$$h_1 \in \arg\min_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \neq h(\vec{x}^i)}.$$

2. *(2 points)* A une itération n quelconque, on a $f_{n-1} = \sum_{k=1}^{n-1} \alpha_k h_k$, les coefficients $(\alpha_k)_{k=1,\ldots n-1}$ et les classifieurs faibles $(h_k)_{k=1,\ldots n-1}$ ayant été fixés aux itérations précédentes. On cherche à incorporer un nouveau classifieur faible h_n dans l'expression de f_n, de sorte que

   $$f_n = f_{n-1} + \alpha_n h_n$$

   Montrer que le classifieur faible h_n permettant de minimiser le risque empirique associé au classifieur f_n est solution de

   $$h_n = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{m} w_i^{(n)} 1_{h_n(\vec{x_i}) = y_i},$$

   où on explicitera la valeur de $w_i^{(n)}$. Que peut-on dire de la valeur de $w_i^{(n)}$ lorsque la donnée $\vec{x^i}$ est correctement classifiée par le classifieur associé à la fonction de décision f_{n-1} ?

   Solution: Le risque empirique associé au classifieur f_n est donné par:

   $$R(f_n) = \sum_{i=1}^{m} \exp\left(-y_i f_{n-1}(\vec{x_i})\right) \exp\left(-y_i \alpha_n h_n(\vec{x_i})\right).$$

   Le classifieur h_n recherché est donc solution de

   $$h_n \in \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{m} w_i^{(n)} \exp\left(-y_i \alpha_n h_n(\vec{x_i})\right),$$

où on a posé $w_i^{(n)} = \exp\left(-y_i f_{n-1}(\vec{x_i})\right)$. On remarque que les minima du problème considéré sont identiques à ceux de

$$h_n \in \arg\min_{h\in\mathcal{F}} \sum_{i=1}^{m} w_i^{(n)} 1_{h_n(\vec{x_i})=y_i},$$

ce qui permet de conclure. Lorsque la donnée $\vec{x^i}$ est correctement classifiée, on a $f_{n-1}(\vec{x^i}) < 0$ et donc $w_i^{(n)} < 1$. Inversement, lorsque la donnée $\vec{x^i}$ est mal classifiée, $w_i^{(n)} > 1$: on donne plus de poids aux données mal classifiées pour entraîner le classifieur faible h_n.

3. *(1 points)* Supposons maintenant le classifieur faible h_n fixé. Comment peut-on sélectionner la valeur du coefficient α_n? Montrer que cela conduit à fixer la valeur de ce paramètre égale à

$$\alpha_n = \log \sqrt{\frac{\sum_{i\in\mathcal{T}} w_i^{(n)}}{\sum_{i\in\mathcal{T}^c} w_i^{(n)}}}$$

où l'ensemble $\mathcal{T}$ est défini par:

$$\mathcal{T} = \{i \in [\![1, m]\!], h_n(\vec{x}^i) = y^i\}.$$

Solution: On fixe la valeur de α_n de sorte que

$$\alpha_n = \arg\min_{\alpha\in\mathbb{R}} J(\alpha) = \sum_{i=1}^{m} w_i^{(n)} \exp(-\alpha h_n(\vec{x_i})).$$

Or, on a

$$J(\alpha) = e^{-\alpha} \sum_{i\in\mathcal{T}} w_i^{(n)} + e^{\alpha} \sum_{i\in\mathcal{T}^c} w_i^{(n)}.$$

Cette quantité est minimale lorsque

$$e^{2\alpha} = \frac{\sum_{i\in\mathcal{T}} w_i^{(n)}}{\sum_{i\in\mathcal{T}^c} w_i^{(n)}},$$

d'où la solution.

4. *(1 points)* Expliquer qualitativement comment on peut fixer en pratique le nombre d'itérations N utilisées pour construire le classifieur f_n.

Solution: En augmentant N, on augmente la complexité du modèle de classification utilisé. Il y a donc un risque de sur-apprentissage. En pratique, on pourra considérer N comme un hyper-paramètre du modèle et fixer sa valeur à l'aide de données de validation.