Correction sommaire de l'examen d'analyse des données - MAT3601 - 2023

11 juin 2023

i) Aucun document autorisé - Admettez le résultat de certaines questions pour passer aux suivantes.

ii) Bien traiter quelques questions rapporte des points, les bâcler toutes n'en rapporte aucun.

iii) Indiquez de manière lisible la question traitée.

*iv) **Soulignez** ou **encadrez** vos résultats.*

v) Écrire votre nom, prénom et numérotez vos copies.

Exercice 1 : ACP

Soit $x_1, \dots, x_n \in \mathbb{R}^d$ nos données, qu'on suppose centrées, et $X \in \mathbb{R}^{n \times d}$ la matrice dont la i-ème ligne contient la i-ème donnée.

1. Donner l'expression de S - la matrice de covariance empirique.

Nous supposons que $d = 4$ et que

$$S = PDP^\top,$$

où $D \in \mathbb{R}^{4\times 4}$ est une matrice diagonale de diagonale $(8, 6, 4, 2)$ et $P \in \mathbb{R}^{4\times 4}$ est une matrice orthogonale. On notera $P = [p_1, p_2, p_3, p_4]$ (p_i est la i-ème colonne de P).

2. Donner les vecteurs propres et les valeurs propres associés de S.
3. Donner l'expression de la 3-ème composante principale.
4. Quelle est la part de la variance expliquée par les 2 premières composantes principales ?
5. Quelle est la dimension de l'espace sélectionné par l'ACP pour que la part de la variance expliquée soit égale à 90% ?

Correction

1. $S = \frac{1}{n} X^\top X$.
2. Comme $S = PDP^\top$ les valeurs propres sont $8, 6, 4, 2$ et les vecteurs propres associés sont p_1, p_2, p_3, p_4. On peut (par exemple) le vérifier par un calcul direct en utilisant l'orthonormalité de $p_1, \dots, p_4$.
3. Par définition, c'est Xp_3 - la projection des données dans l'espace associé vecteur propre associé à la troisième (en ordre décroissant) valeur propre.
4. C'est $\frac{\lambda_1+\lambda_2}{\sum_{i=1}^4 \lambda_i} = \frac{\lambda_1+\lambda_2}{\mathrm{Tr}(S)} = 70\%$.
5. On trouver que $\frac{\lambda_1+\lambda_2+\lambda_3}{\sum_{i=1}^4 \lambda_i} = 90\%$. La dimension de l'espace selectionné est donc de 3.

Exercice 2 : Classification binaire

Nous supposons être en possession d'une suite $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{0, 1\}$ représentant les données étiquetées. On suppose les données i.i.d. de même loi qu'une certaine variable aléatoire $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, notre but est de construire un classifieur relatif à la fonction de perte 0/1.

Nous supposons que $\mathbb{P}(Y = 0) = p_0$, $\mathbb{P}(Y = 1) = p_1$ et nous noterons $g_0 : \mathbb{R}^d \to \mathbb{R}$ (respectivement $g_1 : \mathbb{R}^d \to \mathbb{R}$) la densité conditionnelle de $X|Y = 0$ (respectivement $X|Y = 1$).

1. Soit $x_{n+1} \in \mathbb{R}^d$ une nouvelle donnée. À quelle condition le prédicteur de Bayes prédira 1 comme étiquette associée ?

2. Que se passe-t-il quand p_0 est très proche de 1 ?

   Nous supposons dorénavant qu'il existe $\mu_0, \mu_1 \in \mathbb{R}^d$ t.q. pour $i \in \{0, 1\}$:

$$g_i(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x-\mu_i\|^2} ,$$

3. Pour quels $x \in \mathbb{R}^d$ le prédicteur de Bayes prédira $y = 1$?
4. Pour $d = 2$, $p_0 = p_1 = 0.5$, $\mu_0 = (0, 0)$ et $\mu_1 = (1, 1)$, dessiner la frontière de classification.
5. Comment nomme-t-on ce modèle ?
6. Donner le nom de l'autre modèle de classification binaire vu en cours. Expliquer la différence avec le modèle ci-dessus.

Correction

1. Le prédicteur de Bayes prédira 1 si $\mathbb{P}(Y = 1|X = x_{n+1}) \geqslant \mathbb{P}(Y = 0|X = x_{n+1})$. Dans le cas de l'exercice cette condition équivaut à :

$$p_1 g_1(x_{n+1}) \geqslant p_0 g_0(x_{n+1}) .$$

2. Quand $p_0 \to 1$ on a $p_1 \to 0$ et la condition précédente n'est presque jamais vérifiée. On prédira donc très souvent la valeur 0. C'est logique puisque si p_0 est proche de 1 la plupart des étiquettes seront bien égales à 0.
3. On développe l'expression trouvée dans la question 1 en utilisant les expression de g_0, g_1. En passant au logarithme on trouve :

$$\frac{1}{2}\left(\|x - \mu_0\|^2 - \|x - \mu_1\|^2\right) \geqslant \log\left(\frac{p_0}{p_1}\right) .$$

   En développant les normes on trouve que cette expression se simplifie :

$$\|\mu_0\|^2 - \|\mu_1\|^2 + 2\langle \mu_1 - \mu_0, x\rangle \geqslant 2\log\left(\frac{p_0}{p_1}\right) .$$

   On trouve une expression linéaire ! C'est normal, car ça correspond au cadre du cours où g_0, g_1 sont des densités de loi gaussiennes avec les matrices de covariance égales (identité ici).
4. En utilisant les valeurs données dans l'expression précédente on trouve :

$$2(x_1 + x_2) \geqslant 1 \Leftrightarrow x_2 \geqslant 1/2 - x_1 .$$

   Où $x = (x_1, x_2)$, on trouve l'équation d'une droite (qu'il fallait dessinner).
5. La frontière est linéiare, g_0, g_1 sont des gaussiennes de covariances égales on est dans le cadre du modèle LDA - linear discriminant analysis.
6. L'autre modèle est celui de la QDA - où les matrices de covariances ne sont pas égales, on aurait trouvé une frontière quadratique.

Exercice 3 : Régression linéaire pondérée

Nous allons étudier un problème de régression linéaire pondérée. Soient $y_1, \ldots, y_n \in \mathbb{R}$ et $x_1, \ldots, x_n \in \mathbb{R}$ des données. Le but de la régression linéaire (scalaire) pondérée est de trouver les coefficients $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^2$ t.q.

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} p_i(\alpha + \beta x_i - y_i)^2 , \tag{1}$$

où $p_1, \ldots, p_n$ sont des réels positifs.

1. Pour quelles valeurs des $p_1, \dots, p_n$, les coefficients $\hat{\alpha}, \hat{\beta}$ sont ceux trouvés par la méthode des moindres carrés ordinaire ?
2. Dériver l'équation (1) pour trouver l'équation qui définie $\hat{\alpha}$ en fonction de $\hat{\beta}$ et $y_w = \frac{\sum_{i=1}^n p_i y_i}{\sum_{i=1}^n p_i}$, $x_w = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i}$.
3. En **mettant en évidence les étapes importantes du calcul** montrer que $\hat{\beta}$ vérifie :

$$\sum_{i=1}^n p_i x_i \left(\hat{\beta} (x_i - x_w) - (y_i - y_w) \right) = 0 \ .$$

4. Montrer que pour tout $c \in \mathbb{R}$, $\sum_{i=1}^n p_i c(y_i - y_w) = 0$ en déduire la relation entre

$$\sum_{i=1}^n p_i x_i (y_i - y_w) \quad \text{et} \quad \sum_{i=1}^n p_i (x_i - x_w)(y_i - y_w) \, .$$

5. En déduire l'expression de $\hat{\beta}$ en fonction de $\sum_{i=1}^n p_i (x_i - x_w)^2$ et $\sum_{i=1}^n p_i (x_i - x_w)(y_i - y_w)$.

On peut réécrire l'équation (1) sous forme matricielle :

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\| P(X\beta - Y) \right\|^2 \ , \tag{2}$$

où $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\beta \in \mathbb{R}^d$ et $P \in \mathbb{R}^{n \times n}$ est une matrice diagonale.

6. Préciser la valeur de d et le lien entre Y, X, P, $\tilde{\beta}$ et les paramètres initiaux du problème.

On étudie dorénavant l'équation (2).

7. Supposons que $X^\top P X \in \mathbb{R}^{d \times d}$ est inversible. Exprimer $\tilde{\beta}$ en fonction de P, X, Y.
Indication : on se rappellera que pour tout $z \in \mathbb{R}^d$ et $A \in \mathbb{R}^{d \times d}$ le gradient de la fonction $\theta \mapsto \frac{1}{2} \left\| A\theta - z \right\|^2$ est égal à $A^\top (A\theta - z)$.

On suppose de plus qu'il existe $\beta_* \in \mathbb{R}^d$ t.q.

$$Y = X\beta_* + \xi \, ,$$

où $\xi \in \mathbb{R}^n$ est un vecteur dont les coordonnées $\xi_1, \dots, \xi_n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma_*^2)$, pour $\sigma_* > 0$.

8. Quel type de loi a $\tilde{\beta}$? Quels sont les paramètres qui la caractérise ?
9. Montrer que $\tilde{\beta}$ est un estimateur sans biais de β_*.
10. Calculer la matrice de covariance de $\tilde{\beta}$.
11. **Bonus.** Qu'est-ce qui nous a permis d'affirmer que $\tilde{\beta}$ est unique ? Donner un exemple où il ne l'est pas.

Correction

1. Par exemple pour $p_1 = \ldots = p_n = 1$. Plus généralement à partir du moment où tous les p_i sont égaux et **strictement positifs**.

2. On doit annuler la dérivée par rapport à α en $(\hat{\alpha}, \hat{\beta})$. On dérive l'equation à optimiser par rapport à α et on trouve :

$$2\sum_{i=1}^{n} p_i(\hat{\alpha} + \hat{\beta}x_i - y_i) = 0\,.$$

En regrouppant les termes on trouve

$$\hat{\alpha}\sum_{i=1}^{n} p_i + \hat{\beta}\sum_{i=1}^{n} p_i x_i - \sum_{i=1}^{n} p_i y_i = 0\,.$$

Donc

$$\hat{\alpha} = y_w - \hat{\beta}x_w\,.$$

3. De même on dérive par rapport à β. On trouve :

$$2\sum_{i=1}^{n} p_i x_i(\hat{\alpha} + \hat{\beta}x_i - y_i) = 0\,.$$

On remplace par la valeur trouvée pour $\hat{\alpha}$ dans la question précédente et on trouve :

$$\sum_{i=1}^{n} p_i x_i(y_w - y_i + \hat{\beta}(x_i - x_w)) = 0\,.$$

4. Il suffit de remarquer que $\sum_{i=1}^{n} cp_i y_i = cy_w \sum_{i=1}^{n} p_i$ et que $\sum_{i=1}^{n} cp_i y_w = cy_w \sum_{i=1}^{n} p_i$.
Pour la déduction on remarque x_w est un réel qui **ne dépend pas de** i. Donc en prenant $c = x_w$ on trouve $\sum_{i=1}^{n} p_i x_w(y_i - y_w) = 0$ et

$$\sum_{i=1}^{n} p_i x_i(y_i - y_w) = \sum_{i=1}^{n} p_i(x_i - x_w)(y_i - y_w)\,.$$

5. En utilisant la question 3 on trouve que

$$\hat{\beta} = \frac{\sum_{i=1}^{n} p_i x_i(y_i - y_w)}{\sum_{i=1}^{n} p_i x_i(x_i - x_w)} = \frac{\sum_{i=1}^{n} p_i(x_i - x_w)(y_i - y_w)}{\sum_{i=1}^{n} p_i(x_i - x_w)(x_i - x_w)}\,,$$

où les numérateurs sont égaux par la question précédente et les dénominateurs sont égaux car $\sum_{i=1}^{n} p_i x_w(x_i - x_w) = 0$ (de manière analogue à la question précédente).

6. Comme expliqué dans le cours le $\hat{\alpha}$ est caché dans $\tilde{\beta}$. On trouve $d = 2$, $\tilde{\beta} = (\hat{\alpha}, \hat{\beta})$, $Y \in \mathbb{R}^n$ est le vecteur dont la i-ème coordonnée est égale à y_i, $X \in \mathbb{R}^{n\times d}$ est la matrice dont la i-ème ligne est égale à $(1, x_i)$ et enfin (la partie difficile) P est la matrice diagonale dont la i-ème diagonale est égale à $\sqrt{p_i}$.

7. On doit minimiser $\|P(X\beta - Y)\|^2$ et donc annuler son gradient. On trouve :

$$(PX)^\top(P(X\hat{\beta} - Y)) = 0 \Leftrightarrow X^\top P^\top PX\hat{\beta} = X^\top P^\top PY\,.$$

Comme P est diagonale on a $P^\top = P$. On peut donc simplifier et trouver :

$$X^\top P^2 X\hat{\beta} = X^\top P^2 Y\,.$$

En admetant pour l'instant que $X^\top P^2 X$ est inversible on trouve :

$$\hat{\beta} = (X^\top P^2 X)^{-1} X^\top P^2 Y\,.$$

Preuve que $X^\top P^2 X$ est inversible.

Tout d'abord remarquoqn que $X^\top P^2 X$ est inversible équivaut au fait que $\ker(X^\top P^2 X) = \{0\}$ et donc fait que $\ker PX = \{0\}$. En effet, si $X^\top P^2 X v = 0$ alors $v^\top X^\top P^2 X v = \|PXv\|^2 = 0$ et donc $\ker(X^\top P^2 X) \subset \ker(PX)$. De même, si $PXv = 0$ alors bien sûr $X^\top P^2 X v = 0$ et on a $\ker PX = \ker(X^\top P^2 X)$.
Ainsi comme $X^\top PX$ est inversible on sait que $\ker(P^{1/2}X) = \{0\}$. Maintenant si $v \in \mathbb{R}^d$ est t.q. $PXv = 0$ on a, en notant $y = Xv$,

$$0 = PXv = \begin{pmatrix} \sqrt{p_1} y_1 \\ \sqrt{p_2} y_2 \\ \vdots \\ \sqrt{p_n} y_n \end{pmatrix} = P^{1/2} \begin{pmatrix} p_1^{1/4} y_1 \\ p_2^{1/4} y_2 \\ \vdots \\ p_n^{1/4} y_n \end{pmatrix} = P^{1/2}(P^{1/2} X v)\,.$$

Donc $v \in \ker(PX) \implies v \in \ker(P^1/2X) = \{0\}$. Donc $\ker(PX) = \{0\}$ et $(X^\top P^2 X)$ est bien inversible.

8. ξ est un vecteur gaussien, donc Y l'est aussi (comme transformée affine de ξ) et donc $\tilde{\beta}$ l'est aussi (comme transformée affine de Y). Un vecteur gaussien est determiné par sa moyenne et sa matrice de covariance.
9. On écrit :

$$\tilde{\beta} = (X^\top P^2 X)^{-1} X^\top P^2 Y = (X^\top P^2 X)^{-1} X^\top P^2 (X\beta_* + \xi) = \beta_* + (X^\top P^2 X)^{-1} X^\top P^2 \xi\,. \quad (3)$$

Comme ξ est de moyenne nulle et les autres termes de l'équation sont déterministes on trouve bien :

$$\mathbb{E}[\tilde{\beta}] = \beta_*\,.$$

10. Pour trouver la covariance on utilise la formule $\mathrm{Var}(AZ) = A\,\mathrm{Var}(Z)A^\top$ pour $Z \in \mathbb{R}^n$ un vecteur aléatoire et $A \in \mathbb{R}^{d\times n}$. Ici, en utilisant l'équation (3), on pose $Z = \xi$ et $A = (X^\top P^2 X)^{-1} X^\top P^2$.
11. L'unicité peut être affirmée grace à l'inversibilité de $X^\top P^2 X$ (ou celle de $X^\top PX$). Par exemple si $X = 0$ alors tout vecteur $\beta \in \mathbb{R}^d$ convient.