Examen :
Introduction à l'apprentissage automatique

18 novembre 2022

Aucun document n'est autorisé.

Les questions peuvent être traitées de manière indépendante en admettant les résultats des questions précédentes.

Le barème (sur 20 points, auxquels s'ajoutent 3 points bonus) n'est donné qu'à titre indicatif.

Notations

Dans tout le sujet, on notera :

1. $\mathcal{N}(\mu, \Sigma)$ la loi normale multivariée d'espérance μ et de matrice de variance-covariance Σ (symétrique et semi-définie positive), dont la densité (par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$) lorsque Σ est définie positive est $x \in \mathbb{R}^d \mapsto \frac{1}{\sqrt{2\pi}^d \sqrt{\det(\Sigma)}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$.
2. $\mathcal{B}(p)$ la loi de Bernoulli de paramètre $p \in (0,1)$, qui a pour densité $x \in \{0,1\} \mapsto p^x(1-p)^{1-x}$ (par rapport à la mesure de comptage sur $\{0,1\}$).
3. $\mathcal{B}(m,p)$ la loi binomiale de paramètres $m \in \mathbb{N}^\star$ et $p \in (0,1)$.
4. $\mathbf{1}$ le vecteur rempli de 1 (de taille adéquate).
5. I_n la matrice identité de taille n (la taille peut varier).
6. $\text{sign} : x \in \mathbb{R} \mapsto \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{sinon.} \end{cases}$

Exercice 1 (Modèle mixte, 7 points)

Soit (X,Y) une pair de variables aléatoires à valeurs dans $\mathbb{R}^{d+m} \times \{\pm 1\}$. On souhaite modéliser des données de la forme

$$X_1 = \begin{pmatrix} 0.49 \\ 1.34 \\ -0.70 \\ -1.81 \\ -0.02 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0.69 \\ -0.92 \\ -0.07 \\ -0.82 \\ 0.28 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad X_3 = \begin{pmatrix} -0.74 \\ -1.75 \\ 1.08 \\ -0.15 \\ -0.40 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \dots$$

Pour ce faire, on décompose X en $\begin{bmatrix} U \\ V \end{bmatrix}$ de sorte que (U,V) soit à valeurs dans $\mathbb{R}^d \times \{0,1\}^m$, U représentant les données continues, V les données discrètes. Soit maintenant le modèle

$$\begin{cases} \mathbb{P}(Y=1) = \pi \in (0,1) \\ U|Y=1 \sim \mathcal{N}(\mu_+, \Sigma_+) \text{ est indépendant de } V|Y=1 \sim \mathcal{B}(p_1) \otimes \dots \otimes \mathcal{B}(p_m) \\ U|Y=-1 \sim \mathcal{N}(\mu_-, \Sigma_-) \text{ est indépendant de } V|Y=-1 \sim \mathcal{B}(q_1) \otimes \dots \otimes \mathcal{B}(q_m), \end{cases}$$

avec Σ_+ et Σ_- deux matrices de taille $d \times d$ symétriques et définies positives, $\mu_+, \mu_- \in \mathbb{R}^d$, $p_1, \dots, p_m \in (0,1)$, $q_1, \dots, q_m \in (0,1)$ [1].

1. (1 point) Donner une mesure dominante pour la loi de (X,Y) ainsi qu'une fonction de densité.

2. (1½ points) Montrer que le classifieur de Bayes pour ce modèle est

$$g^\star : (u,v) \in \mathbb{R}^d \times \{0,1\}^m \mapsto \operatorname{sign}\left(\frac{1}{2}u^\top Q u + \alpha^\top u + \beta^\top v + b\right),$$

où

$$\begin{cases} Q &= \Sigma_-^{-1} - \Sigma_+^{-1} \\ \alpha &= \Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_- \\ \beta &= \left[\log\left(\frac{p_1(1-q_1)}{q_1(1-p_1)}\right), \dots, \log\left(\frac{p_m(1-q_m)}{q_m(1-p_m)}\right)\right] \\ b &= \log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log\left(\frac{\det(\Sigma_-)}{\det(\Sigma_+)}\right) + \frac{1}{2}\left(\mu_-^\top\Sigma_-^{-1}\mu_- - \mu_+^\top\Sigma_+^{-1}\mu_+\right) + \sum_{j=1}^m \log\left(\frac{1-p_j}{1-q_j}\right). \end{cases}$$

3. (1 point) On suppose disposer d'un échantillon $(X_1,Y_1), \dots, (X_n,Y_n)$ de même loi que (X,Y). Préciser les estimateurs du maximum de vraisemblance des paramètres $p_1, \dots, p_m$, $q_1, \dots, q_m$.

4. On suppose que $\Sigma_+ = \Sigma_- = I_d$, $\pi = \frac{1}{2}$, $p_1 = \dots = p_m$, $q_1 = \dots = q_m$ et $\mu_+ \neq \mu_-$.
   a) (1 point) Montrer que
   $$\alpha^\top U + b \,\middle|\, Y = 1 \sim \mathcal{N}\left(\delta c, \delta^2\right),$$
   où $\delta = \|\mu_+ - \mu_-\|_{\ell_2}$ et $c = \frac{\delta}{2} + \frac{m}{\delta}\log\left(\frac{1-p_1}{1-q_1}\right)$.

   b) (1½ points) En déduire que $\alpha^\top U + \beta^\top V + b \,|\, Y = 1$ a même loi que
   $$\delta c + \delta A + eB,$$
   où $A \sim \mathcal{N}(0,1) \perp\!\!\!\perp B \sim \mathcal{B}(m, p_1)$ et $e = \log\left(\frac{p_1(1-q_1)}{q_1(1-p_1)}\right)$, puis que
   $$\mathbb{P}(g^\star(X) = -1 \,|\, Y = 1) = \sum_{k=0}^m \binom{m}{k} p_1^k (1-p_1)^{m-k} \Phi\left(-c - \frac{k}{\delta}e\right),$$
   où Φ est la fonction de répartition de $\mathcal{N}(0,1)$.

   c) (1 point) Conclure que lorsque $p_1 = q_1$, $\mathbb{P}(g^\star(X) \neq Y) = \Phi\left(-\frac{\delta}{2}\right)$.

Exercice 2 (Régression à vecteurs supports, 5½ points)

Soient $(X_1,Y_1), \dots, (X_n,Y_n)$ des couples aléatoires i.i.d. à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, $\varepsilon > 0$ et

$$\ell_\varepsilon : u \in \mathbb{R} \mapsto \frac{1}{2}\max(0, |u| - \varepsilon)^2.$$

1. Pour rappel, si $(V_1, \dots, V_m) \sim P_1 \otimes \dots \otimes P_m$ alors les variables aléatoires $V_1, \dots, V_m$ sont indépendantes et $V_i \sim P_i$, $\forall i \in [\![1, m]\!]$.

On considère un RKHS $\mathcal{H}$ de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ et, pour $\lambda > 0$, le problème d'optimisation :

$$\underset{h\in\mathcal{H}}{\text{minimiser}} \quad \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \sum_{i=1}^{n} \ell_\varepsilon\left(Y_i - h(X_i)\right). \tag{P1}$$

1. (1 point) Est-ce un problème de classification ou de régression ? Quelle est sa particularité par rapport à ce qui a été vu en cours ?

2. (1 point) Expliquer pourquoi le problème (P1) est équivalent à

$$\begin{array}{ll} \underset{h\in\mathcal{H},\xi\in\mathbb{R}^n}{\text{minimiser}} & \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \frac{1}{2}\|\xi\|_{\ell_2}^2 \\ \text{s. c.} \quad \forall i \in [\![1,n]\!] & \left\{ \begin{array}{ll} Y_i - h(X_i) \leq \xi_i + \varepsilon & : \alpha_i \geq 0 \\ h(X_i) - Y_i \leq \xi_i + \varepsilon & : \beta_i \geq 0 \\ \xi_i \geq 0 & : \delta_i \geq 0 \end{array} \right. , \end{array} \tag{P2}$$

   où on a donné à titre indicatif les multiplicateurs de Lagrange α_i, β_i et δ_i ($i \in [\![1,n]\!]$) associés à chaque contrainte.

3. ($1\frac{1}{2}$ points) Définir le lagrangien associé à (P2) et énoncer les conditions KKT.

4. (1 point) Montrer qu'en notant $y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, le problème dual à (P2) est

$$\begin{array}{ll} \underset{\alpha\in\mathbb{R}_+^n,\beta\in\mathbb{R}_+^n,\delta\in\mathbb{R}_+^n}{\text{maximiser}} & -\frac{1}{2}\left(\alpha^\top Q\alpha + \beta^\top Q\beta + \|\delta\|_{\ell_2}^2\right) - \alpha^\top P\beta \\ & -\delta^\top(\alpha+\beta) - \alpha^\top(\varepsilon\mathbf{1} - y) - \beta^\top(\varepsilon\mathbf{1} + y), \end{array} \tag{P3}$$

   où

$$\left\{ \begin{array}{ll} K & = (k(X_i, X_j))_{1\leq i,j\leq n} \\ Q & = I_n + \frac{K}{\lambda} \\ P & = I_n - \frac{K}{\lambda}. \end{array} \right.$$

5. (1 point) Soient $\alpha \in \mathbb{R}_+^n, \beta \in \mathbb{R}_+^n$. Montrer que $\forall \delta \in \mathbb{R}_+^n$, $\|\delta\|_{\ell_2}^2 + 2\delta^\top(\alpha+\beta) \geq 0$ et en déduire $\inf_{\delta\in\mathbb{R}_+^n} \|\delta\|_{\ell_2}^2 + 2\delta^\top(\alpha+\beta)$. Montrer que (P3) est équivalent à

$$\underset{\alpha\in\mathbb{R}_+^n,\beta\in\mathbb{R}_+^n}{\text{minimiser}} \quad \frac{1}{2}\alpha^\top Q\alpha + \frac{1}{2}\beta^\top Q\beta + \alpha^\top P\beta + \alpha^\top(\varepsilon\mathbf{1} - y) + \beta^\top(\varepsilon\mathbf{1} + y).$$

6. (1 point (bonus)) On suppose $\varepsilon = 0$ et K inversible. Montrer que (P1) a une unique solution et l'expliciter.

Exercice 3 (Analyse en composantes principales, $7\frac{1}{2}$ points)

Dans cet exercice, pour une matrice notée en majuscule, par exemple $A \in \mathbb{R}^{n\times d}$, nous noterons en minuscule ces colonnes : $a_1, \ldots, a_d \in \mathbb{R}^n$. On rappelle qu'alors

$$\text{range}(A) = \text{span}(\{a_1, \ldots, a_d\}) = \left\{\sum_{i=1}^{d} t_i a_i, t \in \mathbb{R}^d\right\},$$

qui est un sous-espace vectoriel de $\mathbb{R}^n$.

De plus, pour une matrice $Q \in \mathbb{R}^{n\times n}$ réelle symétrique, on appellera décomposition en éléments propres de Q une factorisation $Q = U\Lambda U^\top$, où $U \in \mathbb{R}^{n\times n}$ est une matrice orthogonale ($U^\top U = I_n$), dont les colonnes $u_1, \dots, u_n$ sont les vecteurs propres de Q, et $\Lambda \in \mathbb{R}^{n\times n}$ est une matrice diagonale, dont les éléments diagonaux sont les valeurs propres $\lambda_1 \geq \dots \geq \lambda_n$ de Q rangées par ordre décroissant.

On notera alors, pour tout $p \leq n$, $U_p = [u_1|\dots|u_p] \in \mathbb{R}^{n\times p}$ la matrice rectangulaire des p premières colonnes de U et

$$\Lambda_p = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} \in \mathbb{R}^{p\times p}$$

la matrice diagonale carrée des p premières valeurs propres.

1. a) (1 point) Soient $Q \in \mathbb{R}^{n\times n}$ une matrice symétrique et semi-définie positive de rang r et $x \in \mathbb{R}^n$. Montrer que la projection orthogonale de x sur range(Q), notée Px, vérifie $Px = Q\alpha$ avec $\alpha \in \mathbb{R}^n$ tel que $Qx = Q^2\alpha$.

   b) (1½ points) Soit $Q = U\Lambda U^\top$ une décomposition en éléments propres de Q. Montrer que $Q = U_r\Lambda_r U_r^\top$ puis que le projecteur orthogonal sur range(Q) est $P = U_r U_r^\top$.

2. (1 point) Soit $A \in \mathbb{R}^{n\times d}$ (avec $n \leq d$) une matrice de rang $r \leq n \leq d$. En remarquant que range(A) = range($AA^\top$), déterminer le projecteur orthogonal sur range(A).

3. a) (1 point) Soit $AA^\top = U\Lambda U^\top$ une décomposition en éléments propres de $AA^\top$. On note

$$V = A^\top U_r \Lambda_r^{-1/2}, \qquad \text{avec} \qquad \Lambda_r^{-1/2} = \begin{pmatrix} \lambda_1^{-1/2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_r^{-1/2} \end{pmatrix}.$$

   Montrer que les colonnes de V sont orthonormales.

   b) (1 point) En notant $\Sigma = \Lambda_r^{1/2}$, montrer que $U_r\Sigma V^\top = A$.
   La décomposition de la forme $A = U\Sigma V^\top$, où $U \in \mathbb{R}^{n\times r}$ et $V \in \mathbb{R}^{d\times r}$ sont deux matrices possédant des colonnes orthonormales ($U^\top U = I_r$ et $V^\top V = I_r$) et $\Sigma \in \mathbb{R}^{r\times r}$ est une matrice diagonale, dont les éléments diagonaux $\sigma_1 \geq \dots \geq \sigma_r \geq 0$ sont rangés par ordre décroissant, est appelée décomposition en éléments singuliers (SVD) de A.

4. (1 point) Comment peut-on lier une décomposition en éléments propres de $A^\top A$ à une décomposition en éléments singuliers de A ?

5. (1 point) Soient $\{X_1, \dots, X_n\}$ n vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$ tels que $\mathbb{E}[X_1] = 0$ et $\mathbf{X} \in \mathbb{R}^{n\times d}$ la matrice des données correspondante. Exprimer un estimateur de Var(X_1) en fonction de $\mathbf{X}$ puis décrire une procédure fondée sur la SVD permettant d'implémenter l'analyse en p composantes principales des données, avec $p \leq r$.

6. (1 point (bonus)) Exprimer la matrice (« des données réduites ») de taille $n \times p$ dont la i^e ligne est la compression de X_i en fonction des éléments singuliers déterminés à la question précédente.

7. (1 point (bonus)) En remarquant que, pour tout matrice réelle B, range(B) = $\ker(B^\top)^\perp$ (l'espace orthogonal à $\ker(B^\top)$), montrer que range(B) = range($BB^\top$).

Exercice 1 (Modèle mixte, 7 points)

Soit (X,Y) une pair de variables aléatoires à valeurs dans $\mathbb{R}^{d+m} \times \{\pm 1\}$. On souhaite modéliser des données de la forme

$$X_1 = \begin{pmatrix} 0.49 \\ 1.34 \\ -0.70 \\ -1.81 \\ -0.02 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0.69 \\ -0.92 \\ -0.07 \\ -0.82 \\ 0.28 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad X_3 = \begin{pmatrix} -0.74 \\ -1.75 \\ 1.08 \\ -0.15 \\ -0.40 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \ldots$$

Pour ce faire, on décompose X en $\begin{bmatrix} U \\ V \end{bmatrix}$ de sorte que (U,V) soit à valeurs dans $\mathbb{R}^d \times \{0,1\}^m$, U représentant les données continues, V les données discrètes. Soit maintenant le modèle

$$\begin{cases} \mathbb{P}(Y=1) = \pi \in (0,1) \\ U|Y=1 \sim \mathcal{N}(\mu_+, \Sigma_+) \text{ est indépendant de } V|Y=1 \sim \mathcal{B}(p_1) \otimes \cdots \otimes \mathcal{B}(p_m) \\ U|Y=-1 \sim \mathcal{N}(\mu_-, \Sigma_-) \text{ est indépendant de } V|Y=-1 \sim \mathcal{B}(q_1) \otimes \cdots \otimes \mathcal{B}(q_m), \end{cases}$$

avec Σ_+ et Σ_- deux matrices de taille $d \times d$ symétriques et définies positives, $\mu_+, \mu_- \in \mathbb{R}^d$, $p_1, \ldots, p_m \in (0,1)$, $q_1, \ldots, q_m \in (0,1)$ [1].

1. (1 point) Donner une mesure dominante pour la loi de (X,Y) ainsi qu'une fonction de densité.

---

1) Mesure dominante

Soit Leb est la mesure de Lebesgue sur $\mathbb{R}^d$,
δ_1 est la mesure de comptage sur $\{0,1\}^m$,
δ_2 est la mesure de comptage sur $\{-1,1\}$.

La mesure dominante pour (X,Y) est $\text{Leb} \times \delta_1 \times \delta_2$.

Fonction de densité

soit $X = (U,V)^T$,

$$f_{(X,Y)}(x,y) = \left[\pi \cdot \frac{\exp\left(-\frac{1}{2}(u-\mu_+)^T \Sigma_+^{-1} (u-\mu_+)\right)}{\sqrt{2\pi}^d \sqrt{|\det \Sigma_+|}} \cdot \prod_{i=1}^m p_i^{v_i} \cdot (1-p_i)^{1-v_i}\right]^{\mathbb{1}_{\{Y=1\}}}$$

$$+ \left[(1-\pi) \cdot \frac{\exp\left(-\frac{1}{2}(u-\mu_-)^T \Sigma_-^{-1} (u-\mu_-)\right)}{\sqrt{2\pi}^d \sqrt{|\det \Sigma_-|}} \cdot \prod_{i=1}^m q_i^{v_i} \cdot (1-q_i)^{1-v_i}\right]^{\mathbb{1}_{\{Y=-1\}}}$$

2. (1½ points) Montrer que le classifieur de Bayes pour ce modèle est

$$g^\star : (u,v) \in \mathbb{R}^d \times \{0,1\}^m \mapsto \operatorname{sign}\left(\frac{1}{2}u^\top Q u + \alpha^\top u + \beta^\top v + b\right),$$

où

$$\begin{cases} Q &= \Sigma_-^{-1} - \Sigma_+^{-1} \\ \alpha &= \Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_- \\ \beta &= \left[\log\left(\frac{p_1(1-q_1)}{q_1(1-p_1)}\right), \ldots, \log\left(\frac{p_m(1-q_m)}{q_m(1-p_m)}\right)\right] \\ b &= \log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log\left(\frac{\det(\Sigma_-)}{\det(\Sigma_+)}\right) + \frac{1}{2}\left(\mu_-^\top \Sigma_-^{-1}\mu_- - \mu_+^\top\Sigma_+^{-1}\mu_+\right) + \sum_{j=1}^m \log\left(\frac{1-p_j}{1-q_j}\right). \end{cases}$$

---

(2) D'après la définition,

$$g^*(u,v) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1|X) \geq \mathbb{P}(Y=-1|X) \\ -1 & \text{sinon} \end{cases}$$

$$= \operatorname{sign}\left(\log \frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=-1|X)}\right)$$

$$= \operatorname{sign}\left(\log \frac{\mathbb{P}(X,Y=1)}{\mathbb{P}(X,Y=-1)}\right)$$

$$\log \mathbb{P}(X,Y=1) = \log \pi - \frac{1}{2}(u-\mu_+)^T\Sigma_+^{-1}(u-\mu_+) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(\det|\Sigma_+|) + \sum_{i=1}^m v_i \cdot \log p_i + \sum_{i=1}^m (1-v_i)\cdot \log(1-p_i)$$

$$\log \mathbb{P}(X,Y=-1) = \log(1-\pi) - \frac{1}{2}(u-\mu_-)^T\Sigma_-^{-1}(u-\mu_-) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(\det|\Sigma_-|) + \sum_{i=1}^m v_i \cdot \log q_i + \sum_{i=1}^m (1-v_i)\cdot \log(1-q_i)$$

alors

$$\log \frac{\mathbb{P}(X,Y=1)}{\mathbb{P}(X,Y=-1)} = \log \mathbb{P}(X,Y=1) - \log \mathbb{P}(X,Y=-1)$$

$$= \frac{1}{2}u^TQu + \alpha^T u + \beta^T v + b$$

3. (1 point) On suppose disposer d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y). Préciser les estimateurs du maximum de vraisemblance des paramètres $p_1, \ldots, p_m$, $q_1, \ldots, q_m$.

(3) D'après (2)

$$\log f_{(X,Y)}(x,y) = \mathbb{1}_{\{Y=1\}} \cdot \log \mathbb{P}(X, Y=1) + \mathbb{1}_{\{Y=-1\}} \cdot \log \mathbb{P}(X, Y=-1)$$

alors

$$\ell(p,q) = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k=1\}} \cdot \log \mathbb{P}(X_k, Y_k=1) + \sum_{k=1}^{n} \mathbb{1}_{\{Y_k=-1\}} \cdot \log \mathbb{P}(X_k, Y_k=-1)$$

on a

① $$\frac{\partial \ell}{\partial p_i}(p,q) = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k=1\}} \cdot \left(\frac{V_{ki}}{p_i} + \frac{1-V_{ki}}{1-p_i}\right)$$

pour $\frac{\partial \ell}{\partial p_i} = 0$, on a $$\hat{p}_i = \frac{\sum_{k=1}^{n} \mathbb{1}_{\{Y_k=1\}} \cdot V_{ki}}{\sum_{k=1}^{n} \mathbb{1}_{\{Y_k=1\}} \cdot (2V_{ki}-1)}$$

② $$\frac{\partial \ell}{\partial q_i}(p,q) = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k=-1\}} \cdot \left(\frac{V_{ki}}{q_i} + \frac{1-V_{ki}}{1-q_i}\right)$$

pour $\frac{\partial \ell}{\partial q_i} = 0$, on a $$\hat{q}_i = \frac{\sum_{k=1}^{n} \mathbb{1}_{\{Y_k=-1\}} \cdot V_{ki}}{\sum_{k=1}^{n} \mathbb{1}_{\{Y_k=-1\}} \cdot (2V_{ki}-1)}$$

4. On suppose que $\Sigma_+ = \Sigma_- = I_d$, $\pi = \frac{1}{2}$, $p_1 = \cdots = p_m$, $q_1 = \cdots = q_m$ et $\mu_+ \neq \mu_-$.

a) (1 point) Montrer que

$$\alpha^\top U + b \,\big|\, Y = 1 \sim \mathcal{N}\left(\delta c, \delta^2\right),$$

où $\delta = \|\mu_+ - \mu_-\|_{\ell_2}$ et $c = \frac{\delta}{2} + \frac{m}{\delta} \log\left(\frac{1-p_1}{1-q_1}\right)$.

(4) (a) $U \mid Y=1 \sim N(\mu_+, \Sigma_+)$

$\alpha^T U \mid Y=1 \sim N(\alpha^T \mu_+, \alpha \Sigma_+ \alpha^T)$

$\alpha^T U + b \mid Y=1 \sim N(\alpha^T \mu_+ + b, \alpha \Sigma_+ \alpha^T)$

$$\begin{aligned} \alpha^T \mu_+ + b &= \mu_+^T \mu_+ - \mu_-^T \mu_+ + \frac{1}{2} \mu_-^T \mu_- - \frac{1}{2} \mu_+^T \mu_+ + m \cdot \log \frac{1-p_1}{1-q_1} \\ &= \frac{1}{2} \|\mu_+ - \mu_-\|_{\ell_2}^2 + m \cdot \log \frac{1-p_1}{1-q_1} \\ &= \delta \cdot c \end{aligned}$$

$$\alpha \Sigma_+ \alpha^T = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T = \|\mu_+ - \mu_-\|_{\ell_2}^2 = \delta^2$$

b) (1½ points) En déduire que $\alpha^\top U + \beta^\top V + b \,|\, Y = 1$ a même loi que

$$\delta c + \delta A + eB,$$

où $A \sim \mathcal{N}(0,1) \perp\!\!\!\perp B \sim \mathcal{B}(m, p_1)$ et $e = \log\left(\frac{p_1(1-q_1)}{q_1(1-p_1)}\right)$, puis que

$$\mathbb{P}(g^\star(X) = -1 \,|\, Y = 1) = \sum_{k=0}^{m} \binom{m}{k} p_1^k (1-p_1)^{m-k} \Phi\left(-c - \frac{k}{\delta} e\right),$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$.

(b) D'après (a), $\alpha^T U + b \,|\, Y=1 \sim \delta c + \delta A$

et

$$V \,|\, Y=1 \sim \mathcal{B}(m, p_1)$$
$$\beta^T V \,|\, Y=1 \sim eB$$

comme $U|Y=1 \perp\!\!\!\perp V|Y=1$, alors

$$\alpha^T U + \beta^T V + b \sim \delta c + \delta A + eB$$

Dans la condition de (3),

$$g^*(X) = g^*(U,V) = \text{sign}(\alpha^T U + \beta^T V + b)$$

donc

$$\begin{aligned}
\mathbb{P}(g^*(X) = -1 \,|\, Y=1) &= \mathbb{P}(\alpha^T U + \beta^T V + b < 0 \,|\, Y=1) \\
&= \mathbb{P}(\delta c + \delta A + eB < 0) \\
&= \mathbb{E}\left[\mathbb{1}_{\{\delta c + \delta A + eB < 0\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{A < -c - \frac{e}{\delta} B\}} \,\middle|\, B\right]\right] \\
&= \mathbb{E}\left[\Phi(-c - \tfrac{e}{\delta} B)\right] \\
&= \sum_{k=0}^{m} \binom{m}{k} p_1^k \cdot (1-p_1)^{m-k} \cdot \Phi(-c - \tfrac{e}{\delta} k)
\end{aligned}$$

c) (1 point) Conclure que lorsque $p_1 = q_1$, $\mathbb{P}(g^\star(X) \neq Y) = \Phi\left(-\frac{\delta}{2}\right)$.

(c) Lorsque $p_1 = q_1$, $e = 0$, $c = \frac{\delta}{2}$

$$\mathbb{P}(g^*(X) = -1 \mid Y = 1) = \mathbb{E}[\Phi(-c)]$$
$$= \Phi(-\tfrac{\delta}{2})$$

la même

$$\alpha^T U + b \mid Y = -1 \sim -\delta c + \delta A$$
$$\beta^T V \mid Y = -1 \sim eB$$

$$\mathbb{P}(g^*(X) = 1 \mid Y = -1) = \mathbb{P}(-\delta c + \delta A + eB > 0)$$
$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{A > c - \tfrac{e}{\delta}B\} \mid B\right]\right]$$
$$= \mathbb{E}\left[1 - \Phi(c - \tfrac{e}{\delta}B)\right]$$
$$= 1 - \Phi(\tfrac{\delta}{2})$$
$$= \Phi(-\tfrac{\delta}{2})$$

alors

$$\mathbb{P}(g^*(X) \neq Y) = \mathbb{P}(g^*(X) = 1 \mid Y = -1) \cdot \mathbb{P}(Y = -1)$$
$$+ \mathbb{P}(g^*(X) = -1 \mid Y = 1) \cdot \mathbb{P}(Y = 1)$$
$$= \Phi(-\tfrac{\delta}{2}) \cdot \tfrac{1}{2} + \Phi(-\tfrac{\delta}{2}) \cdot \tfrac{1}{2}$$
$$= \Phi(-\tfrac{\delta}{2})$$

Exercice 2 (Régression à vecteurs supports, 5½ points)

Soient $(X_1, Y_1), \dots, (X_n, Y_n)$ des couples aléatoires i.i.d. à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, $\varepsilon > 0$ et

$$\ell_\varepsilon : u \in \mathbb{R} \mapsto \frac{1}{2} \max(0, |u| - \varepsilon)^2.$$

On considère un RKHS $\mathcal{H}$ de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ et, pour $\lambda > 0$, le problème d'optimisation :

$$\underset{h\in\mathcal{H}}{\text{minimiser}} \ \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \sum_{i=1}^n \ell_\varepsilon (Y_i - h(X_i)) . \tag{P1}$$

1. (1 point) Est-ce un problème de classification ou de régression ? Quelle est sa particularité par rapport à ce qui a été vu en cours ?

(1) C'est un problème de régression.

Dans le cours, la fonction de loss est

$\ell_\varepsilon^{(c)} : u \in \mathbb{R} \longrightarrow \max(0, |u| - \varepsilon)$

qui est different de ℓ_ε.

2. (1 point) Expliquer pourquoi le problème (P1) est équivalent à

$$\underset{h\in\mathcal{H}, \xi\in\mathbb{R}^n}{\text{minimiser}} \ \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{2} \|\xi\|_{\ell_2}^2$$
$$\text{s. c.} \quad \forall i \in [\![1, n]\!] \begin{cases} Y_i - h(X_i) \le \xi_i + \varepsilon & : \alpha_i \ge 0 \\ h(X_i) - Y_i \le \xi_i + \varepsilon & : \beta_i \ge 0 \\ \xi_i \ge 0 & : \delta_i \ge 0 \end{cases}, \tag{P2}$$

où on a donné à titre indicatif les multiplicateurs de Lagrange α_i, β_i et δ_i ($i \in [\![1, n]\!]$) associés à chaque contrainte.

(2) $\forall u \in \mathbb{R}$, $\ell_\varepsilon(u) = \inf\limits_{\substack{\xi \ge 0 \\ \xi \ge |u| - \varepsilon}} \frac{1}{2}\xi^2 = \inf\limits_{\substack{u \le \xi + \varepsilon \\ -u \le \xi + \varepsilon \\ \xi \ge 0}} \frac{1}{2}\xi^2$

alors pour $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$

$$\underset{h\in H}{\text{minimiser}} \ \frac{\lambda}{2}\|h\|_{H}^2 + \sum_{i=1}^n \ell_\varepsilon(Y_i - h(X_i))$$

$\Longleftrightarrow$

$$\underset{\substack{h\in H \\ \xi \in \mathbb{R}^n}}{\text{minimiser}} \ \frac{\lambda}{2}\|h\|_{H}^2 + \frac{1}{2}\sum_{i=1}^n \xi_i^2$$

$$\text{s.c.} \quad \forall i \in [\![1,n]\!] \quad \begin{cases} Y_i - h(X_i) \leq \xi_i + \varepsilon \\ h(X_i) - Y_i \leq \xi_i + \varepsilon \\ \xi_i \geq 0 \end{cases}$$

3. ($1\frac{1}{2}$ points) Définir le lagrangien associé à (P2) et énoncer les conditions KKT.

(3) D'après (2), soit $\forall i \in [\![1,n]\!]$

$$f(h,\xi) = \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \frac{1}{2}\|\xi\|_{\ell_2}^2$$

$$g_i^{(1)}(h,\xi) = Y_i - h(X_i) - \xi_i - \varepsilon$$

$$g_i^{(2)}(h,\xi) = h(X_i) - Y_i - \xi_i - \varepsilon$$

$$g_i^{(3)}(h,\xi) = -\xi_i$$

alors on a

$$\forall (h,\xi,\alpha,\beta,\delta) \in \mathcal{H} \times \mathbb{R}^n \times \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{R}_+^n$$

$$\begin{aligned} \mathcal{L}(h,\xi,\alpha,\beta,\delta) &= f(h,\xi) + \sum_{i=1}^n \alpha_i \cdot g_i^{(1)}(h,\xi) + \sum_{i=1}^n \beta_i \cdot g_i^{(2)}(h,\xi) + \sum_{i=1}^n \delta_i \cdot g_i^{(3)}(h,\xi) \\ &= \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \frac{1}{2}\|\xi\|_{\ell_2}^2 \\ &+ \sum_{i=1}^n \alpha_i \cdot (Y_i - h(X_i) - \xi_i - \varepsilon) \\ &+ \sum_{i=1}^n \beta_i \cdot (h(X_i) - Y_i - \xi_i - \varepsilon) \\ &- \sum_{i=1}^n \delta_i \cdot \xi_i \end{aligned}$$

Le condition KKT

Si le problème (P_2) est convexe et si les conditions de Slater sont vérifiées.

Alors (h_n^*, ξ_n^*) est la solution de (P_2) et $(\alpha^*, \beta^*, \delta^*)$ est la solution du probléme dual, ssi

① $g_i^{(k)}(h_n^*, \xi_n) \leqslant 0$, $\forall i \in [\![1,n]\!]$, $\forall k \in [\![1,3]\!]$

② $\alpha^* \geqslant 0$, $\beta^* \geqslant 0$, $\delta^* \geqslant 0$

③ $\forall i \in [\![1,n]\!]$, $\alpha_i^* = 0$ ou $g_i^{(1)}(h_n^*, \xi_n^*) = 0$

$\beta_i^* = 0$ ou $g_i^{(2)}(h_n^*, \xi_n^*) = 0$

$\delta_i^* = 0$ ou $g_i^{(3)}(h_n^*, \xi_n^*) = 0$

④ $\nabla_h \mathcal{L}(h, \xi, \alpha, \beta, \delta) = 0$

$\nabla_\xi \mathcal{L}(h, \xi, \alpha, \beta, \delta) = 0$

$$\Longrightarrow \begin{cases} h_n^* = \frac{1}{\lambda} \cdot \sum_{i=1}^{n} (\alpha_i - \beta_i) \cdot k(\cdot, X_i) \\ \xi_n^* = \alpha + \beta + \delta \end{cases}$$

日期: /

Le condition KKT

Si h_n solution de (P_1), alors (D) a la solution $(\alpha^*, \beta^*, \delta^*)$ tel que

- $h_n = \frac{1}{\lambda} \cdot \sum_{i=1}^{n} (\alpha_i^* - \beta_i^*) \cdot k(\cdot, X_i)$
- $\forall i \in [\![1, n]\!]$, $Y_i - h_n(X_i) > \varepsilon \implies \alpha_i^* = 0$

$$Y_i - h_n(X_i) < \varepsilon \implies \alpha_i^* = 1$$

$$h_n(X_i) - Y_i > \varepsilon \implies \beta_i^* = 0$$

$$h_n(X_i) - Y_i < \varepsilon \implies \beta_i^* = 1$$

$$h_n^* = \frac{1}{\lambda} \cdot \sum_{i=1}^{n} (\alpha_i - \beta_i) \cdot k(\cdot, X_i)$$

$$\xi_n^* = \alpha + \beta + \delta$$

$$\begin{aligned} L(h, \xi, \alpha, \beta, \delta) = & \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{2} \|\xi\|_{\ell_2}^2 \\ & + \sum_{i=1}^{n} \alpha_i \cdot (Y_i - h(X_i) - \xi_i - \varepsilon) \\ & + \sum_{i=1}^{n} \beta_i \cdot (h(X_i) - Y_i - \xi_i - \varepsilon) \\ & - \sum_{i=1}^{n} \delta_i \cdot \xi_i \end{aligned}$$

4. (1 point) Montrer qu'en notant $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, le problème dual à (P2) est

$$\underset{\alpha\in\mathbb{R}^n_+,\beta\in\mathbb{R}^n_+,\delta\in\mathbb{R}^n_+}{\text{maximiser}} \quad -\frac{1}{2}\left(\alpha^\top Q\alpha + \beta^\top Q\beta + \|\delta\|^2_{\ell_2}\right) - \alpha^\top P\beta$$
$$-\delta^\top(\alpha+\beta) - \alpha^\top(\varepsilon\mathbf{1} - y) - \beta^\top(\varepsilon\mathbf{1} + y), \tag{P3}$$

où

$$\begin{cases} K &= (k(X_i, X_j))_{1\le i,j\le n} \\ Q &= I_n + \frac{K}{\lambda} \\ P &= I_n - \frac{K}{\lambda}. \end{cases}$$

(4) D'après (3)

$$D(\alpha,\beta,\delta) = \inf_{\substack{h\in\mathcal{H} \\ \xi\in\mathbb{R}^n}} \mathcal{L}(h,\xi,\alpha,\beta,\delta) = \mathcal{L}(h_n^*, b_n^*, \alpha,\beta,\delta)$$

$$= \frac{\lambda}{2}\cdot\left\|\frac{1}{\lambda}\sum_{i=1}^n(\alpha_i-\beta_i)\cdot k(\cdot,X_i)\right\|^2_{\mathcal{H}} + \frac{1}{2}\|\alpha+\beta+\delta\|^2_{\ell_2}$$
$$-\alpha^T(\varepsilon\mathbb{1}-y) - \alpha^T(\alpha+\beta+\delta) - \sum_{i=1}^n\alpha_i\frac{1}{\lambda}\sum_{j=1}^n(\alpha_j-\beta_j)\cdot k(X_i,X_j)$$
$$-\beta^T(\varepsilon\mathbb{1}-y) - \beta^T(\alpha+\beta+\delta) - \sum_{i=1}^n\beta_i\cdot\frac{1}{\lambda}\sum_{j=1}^n(\alpha_j-\beta_j)\cdot k(X_i,X_j)$$
$$-\delta^T(\alpha+\beta+\delta)$$

$$= \frac{\lambda}{2}\cdot\frac{1}{\lambda^2}\cdot\sum_{i=1}^n\sum_{j=1}^n(\alpha_i-\beta_i)(\alpha_j-\beta_j)\cdot k(X_i,X_j) - \frac{1}{2}\|\alpha+\beta+\delta\|^2_{\ell_2}$$
$$-\alpha^T(\varepsilon\mathbb{1}-y) - \beta^T(\varepsilon\mathbb{1}-y) - \alpha^T K(\alpha-\beta) - \beta^T K(\alpha-\beta)$$

$$= \frac{1}{2}(\alpha-\beta)^T\frac{K}{\lambda}(\alpha-\beta) - \alpha^T\frac{K}{\lambda}(\alpha-\beta) - \beta^T\frac{K}{\lambda}(\alpha-\beta)$$
$$-\alpha^T(\varepsilon\mathbb{1}-y) - \beta^T(\varepsilon\mathbb{1}-y) - \frac{1}{2}\|\alpha+\beta+\delta\|^2_{\ell_2}$$

$$= -\frac{1}{2}\alpha^T\frac{K}{\lambda}\alpha - \frac{1}{2}\beta^T\frac{K}{\lambda}\beta - \alpha^T\frac{K}{\lambda}\beta$$
$$-\alpha^T(\varepsilon\mathbb{1}-y) - \beta^T(\varepsilon\mathbb{1}-y)$$
$$-\frac{1}{2}(\alpha^T\alpha + \beta^T\beta + \delta^T\delta + 2\alpha^T\beta + 2\delta^T\alpha + 2\delta^T\beta)$$

日期:　　　/

$$= -\frac{1}{2}\alpha^T Q\alpha - \frac{1}{2}\beta^T Q\beta - \alpha^T P\beta - \frac{1}{2}\|\delta\|_{\ell_2}^2 - \delta^T(\alpha+\beta)$$
$$- \alpha^T(\varepsilon\mathbb{1}-y) - \beta^T(\varepsilon\mathbb{1}-y)$$

5. (1 point) Soient $\alpha \in \mathbb{R}_+^n, \beta \in \mathbb{R}_+^n$. Montrer que $\forall \delta \in \mathbb{R}_+^n$, $\|\delta\|_{\ell_2}^2 + 2\delta^\top(\alpha+\beta) \geq 0$ et en déduire $\inf_{\delta \in \mathbb{R}_+^n} \|\delta\|_{\ell_2}^2 + 2\delta^\top(\alpha+\beta)$. Montrer que (P3) est équivalent à

$$\underset{\alpha\in\mathbb{R}_+^n, \beta\in\mathbb{R}_+^n}{\text{minimiser}} \ \frac{1}{2}\alpha^\top Q\alpha + \frac{1}{2}\beta^\top Q\beta + \alpha^\top P\beta + \alpha^\top(\varepsilon\mathbf{1}-y) + \beta^\top(\varepsilon\mathbf{1}+y).$$

(5) $\|\delta\|_{\ell_2}^2 + 2\delta^T(\alpha+\beta) = \|\alpha+\beta+\delta\|_{\ell_2}^2 - \|\alpha+\beta\|_{\ell_2}^2$

$\geqslant 0$

quand $\delta = 0$, on a

$$\inf_{\delta\in\mathbb{R}_+^n} \|\delta\|_{\ell_2}^2 + 2\delta^T(\alpha+\beta) = 0$$

alors (P3)

$$\Longleftrightarrow \underset{\alpha,\beta,\delta}{\text{minimiser}} \ \frac{1}{2}(\alpha^T Q\alpha + \beta^T Q\beta + \|\delta\|_{\ell_2}^2) + \alpha^T P\beta + \delta^T(\alpha+\beta)$$
$$+ \alpha^T(\varepsilon\mathbb{1}-y) + \beta^T(\varepsilon\mathbb{1}-y)$$

$$\Longleftrightarrow \underset{\alpha,\beta}{\text{minimiser}} \ \frac{1}{2}(\alpha^T Q\alpha + \beta^T Q\beta) + \alpha^T P\beta + \alpha^T(\varepsilon\mathbb{1}-y) + \beta^T(\varepsilon\mathbb{1}-y)$$

6. (1 point (bonus)) On suppose $\varepsilon = 0$ et K inversible. Montrer que (P1) a une unique solution et l'expliciter.

(6) Pour $\varepsilon = 0$ et k inversible, on a

$$(P4) \quad \underset{\alpha \in \mathbb{R}^n_+, \beta \in \mathbb{R}^n_+}{\text{minimiser}} \ \frac{1}{2}\alpha^T Q \alpha + \frac{1}{2}\beta^T Q \beta + \alpha^T P \beta - \alpha^T y + \beta^T y$$

$$\nabla_\alpha F(\alpha, \beta) = Q\alpha + P\beta - y$$

$$\nabla_\beta F(\alpha, \beta) = Q\beta + P\alpha + y$$

on prend $\begin{cases} Q\alpha + P\beta - y = 0 \\ Q\beta + P\alpha + y = 0 \end{cases}$

alors $\alpha^* = \frac{y}{2}$ $\quad \beta^* = -\frac{y}{2}$

on a la unique solution

$$h_n^* = \frac{1}{\lambda} \sum_{i=1}^{n} Y_i \cdot k(\cdot, X_i)$$

Exercice 3 (Analyse en composantes principales, $7\frac{1}{2}$ points)

Dans cet exercice, pour une matrice notée en majuscule, par exemple $A \in \mathbb{R}^{n\times d}$, nous noterons en minuscule ces colonnes : $a_1, \dots, a_d \in \mathbb{R}^n$. On rappelle qu'alors

$$\mathrm{range}(A) = \mathrm{span}(\{a_1, \dots, a_d\}) = \left\{ \sum_{i=1}^{d} t_i a_i, t \in \mathbb{R}^d \right\},$$

qui est un sous-espace vectoriel de $\mathbb{R}^n$.

De plus, pour une matrice $Q \in \mathbb{R}^{n\times n}$ réelle symétrique, on appellera décomposition en éléments propres de Q une factorisation $Q = U\Lambda U^\top$, où $U \in \mathbb{R}^{n\times n}$ est une matrice orthogonale ($U^\top U = I_n$), dont les colonnes $u_1, \dots, u_n$ sont les vecteurs propres de Q, et $\Lambda \in \mathbb{R}^{n\times n}$ est une matrice diagonale, dont les éléments diagonaux sont les valeurs propres $\lambda_1 \geq \dots \geq \lambda_n$ de Q rangées par ordre décroissant.

On notera alors, pour tout $p \leq n$, $U_p = [u_1|\dots|u_p] \in \mathbb{R}^{n\times p}$ la matrice rectangulaire des p premières colonnes de U et

前p个列向量

$$\Lambda_p = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} \in \mathbb{R}^{p\times p}$$

la matrice diagonale carrée des p premières valeurs propres.

1. a) (1 point) Soient $Q \in \mathbb{R}^{n\times n}$ une matrice symétrique et semi-définie positive de rang r et $x \in \mathbb{R}^n$. Montrer que la projection orthogonale de x sur range(Q), notée Px, vérifie $Px = Q\alpha$ avec $\alpha \in \mathbb{R}^n$ tel que $Qx = Q^2\alpha$.

---

1) (a) Pour $Q = (q_1, \dots, q_n) \in \mathbb{R}^{n\times n}$, on a

$\mathrm{rang}(Q) = \mathrm{span}(\{q_1, \dots, q_n\}) = \left\{ \sum_{i=1}^{n} \alpha_i \cdot q_i, \alpha \in \mathbb{R}^d \right\}$

Si Px est une projection orthogonale de x sur rang(Q), alors $Px \in \left\{ \sum_{i=1}^{n} \alpha_i q_i, \alpha \in \mathbb{R}^n \right\}$

c-à-d il existe $\alpha \in \mathbb{R}^n$, tel que

$Px = Q\alpha$

Et comme $x \in \mathbb{R}^n$, $Qx = \sum_{i=1}^{n} x_i q_i \in \mathrm{rang}(Q)$, alors

$Qx = P(Qx) = QPx = QQ\alpha = Q^2\alpha$

投影性质？

b) (1½ points) Soit $Q = U\Lambda U^\top$ une décomposition en éléments propres de Q. Montrer que $Q = U_r\Lambda_r U_r^\top$ puis que le projecteur orthogonal sur range(Q) est $P = U_r U_r^\top$.

(b) Comme $Q \in \mathbb{R}^{n\times n}$ est symétrique et de rang r,
alors les valeurs propres sont $\lambda_1 \geqslant \cdots \geqslant \lambda_r$ et
les autres sont 0.
Donc $Q = U\Lambda U^T = U_r \Lambda_r U_r^T$
D'après (a)
$Px = Q\alpha = Q^{-1}QQ\alpha = Q^{-1}Qx$
donc $P = Q^{-1}Q = U_r\Lambda_r^{-1}U_r^T U_r \Lambda_r U_r^T = U_r U_r^T$

2. (1 point) Soit $A \in \mathbb{R}^{n\times d}$ (avec $n \leq d$) une matrice de rang $r \leq n \leq d$. En remarquant que range(A) = range($AA^\top$), déterminer le projecteur orthogonal sur range(A).

(2) comme $A \in \mathbb{R}^{n\times d}$, alors $AA^T \in \mathbb{R}^{n\times n}$
et A est de rang r, donc AA^T est de rang r
il existe $U_r \in \mathbb{R}^{r\times r}$, $\Lambda_r \in \mathbb{R}^{r\times r}$, tel que

$$AA^T = U_r \Lambda_r U_r^T$$

d'après (1), le projecteur orthogonal sur rang(AA^T) = rang(A)
est $P = U_r U_r^T$

3. a) (1 point) Soit $AA^\top = U\Lambda U^\top$ une décomposition en éléments propres de $AA^\top$. On note

$$V = A^\top U_r \Lambda_r^{-1/2}, \qquad \text{avec} \qquad \Lambda_r^{-1/2} = \begin{pmatrix} \lambda_1^{-1/2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_r^{-1/2} \end{pmatrix}.$$

Montrer que les colonnes de V sont orthonormales.

(3) (a)

$$\begin{aligned} V^\top V &= \Lambda_r^{-\frac{1}{2}} U_r^\top A A^\top U_r \Lambda_r^{-\frac{1}{2}} \\ &= \Lambda_r^{-\frac{1}{2}} U_r^\top U_r \Lambda_r U_r^\top U_r \Lambda_r^{-\frac{1}{2}} \\ &= \Lambda_r^{-\frac{1}{2}} \Lambda_r \Lambda_r^{-\frac{1}{2}} \\ &= I_n \end{aligned}$$

donc les colonnes de V sont orthonormales

b) (1 point) En notant $\Sigma = \Lambda_r^{1/2}$, montrer que $U_r \Sigma V^\top = A$.
La décomposition de la forme $A = U\Sigma V^\top$, où $U \in \mathbb{R}^{n\times r}$ et $V \in \mathbb{R}^{d\times r}$ sont deux matrices possédant des colonnes orthonormales ($U^\top U = I_r$ et $V^\top V = I_r$) et $\Sigma \in \mathbb{R}^{r\times r}$ est une matrice diagonale, dont les éléments diagonaux $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ sont rangés par ordre décroissant, est appelée décomposition en éléments singuliers (SVD) de A.

(3) (b)

$$\begin{aligned} U_r \Sigma V^\top &= U_r \Lambda_r^{\frac{1}{2}} \Lambda_r^{-\frac{1}{2}} U_r^\top A \\ &= I_n A \\ &= A \end{aligned}$$

4. (1 point) Comment peut-on lier une décomposition en éléments propres de $A^\top A$ à une décomposition en éléments singuliers de A ?

(4) Pour la décomposition en éléments propres de AA^T,

$$AA^T = U_r \Lambda_r U_r^T$$

pour la décomposition en éléments singuliers de A,

$$A = U_r \Sigma V^T$$

avec $\begin{cases} \Sigma = \Lambda_r^{\frac{1}{2}} \\ V = A^T U_r \Lambda_r^{-\frac{1}{2}} \end{cases}$

Alors, pour la décomposition en éléments propres de A^TA,

$$A^TA = V \Sigma U_r^T U_r \Sigma V^T$$
$$= V \Lambda_r V^T$$

donc $V = (v_1, \dots, v_r)$ sont des vecteurs propres de A^TA

5. (1 point) Soient $\{X_1, \dots, X_n\}$ n vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$ tels que $\mathbb{E}[X_1] = 0$ et $\mathbf{X} \in \mathbb{R}^{n\times d}$ la matrice des données correspondante. Exprimer un estimateur de $\mathrm{Var}(X_1)$ en fonction de $\mathbf{X}$ puis décrire une procédure fondée sur la SVD permettant d'implémenter l'analyse en p composantes principales des données, avec $p \leq r$.

(5) D'après la définition $Var(X_1) = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])^T\right]$

donc l'estimateur est

$$\mathbb{V}(\mathbb{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(X_i - \frac{1}{n}\sum_{k=1}^{n}X_k\right)^T$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_iX_i^T - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}X_iX_k^T - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}X_jX_i^T + \frac{1}{n^2}\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{k=1}^{n}X_k^T\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_iX_i^T - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j^T$$

$$= \frac{1}{n}\mathbb{X}^T\mathbb{X} - \frac{1}{n^2}(\mathbb{X}^T\mathbb{1})^T(\mathbb{X}^T\mathbb{1})$$

- Le PCA probléme est équivalent à

$$\underset{\substack{U\in\mathbb{R}^{d\times p}\\ U^TU = I_p}}{\text{maximize}}\ tr(U^T\mathbb{V}(X)U)$$

- <u>L'analyse en p composantes principales</u> (SVD)

Input : $\mathbb{X} \in \mathbb{R}^{n\times d}$, p

$\sigma_1, \dots, \sigma_p \longleftarrow$ p leading singular values of $\mathbb{X}$

$V \longleftarrow$ p leading left singular vectors of $\mathbb{X}$

$U \longleftarrow [\sigma_1 V_1, \sigma_2 V_2, \dots, \sigma_p V_p]$

Output : $U \in \mathbb{R}^{n\times p}$

6. (1 point (bonus)) Exprimer la matrice (« des données réduites ») de taille $n \times p$ dont la i^{e} ligne est la compression de X_i en fonction des éléments singuliers déterminés à la question précédente.

(6)

7. (1 point (bonus)) En remarquant que, pour tout matrice réelle B, $\mathrm{range}(B) = \ker(B^\top)^\perp$ (l'espace orthogonal à $\ker(B^\top)$), montrer que $\mathrm{range}(B) = \mathrm{range}(BB^\top)$.

(7) Soit $B \in \mathbb{R}^{n\times m}$, $\ker(B^T) = \{x \in \mathbb{R}^n \mid B^T x = 0\}$

$\mathrm{range}(B) = \{\sum_{i=1}^{m} b_i \cdot \alpha_i \mid \alpha \in \mathbb{R}^m\} = \{x \in \mathbb{R}^n \mid B^T x \neq 0\}$

① Pour montrer $\mathrm{range}(B) \subseteq \mathrm{range}(BB^T)$

il suffit de montrer que $\forall\, y \in \mathrm{range}(B)$, $y \in \mathrm{range}(BB^T)$

$\forall\, y \in \mathrm{range}(B)$, il existe $x \in \mathbb{R}^m$

$y = Bx$

il suffit de trouver $z \in \mathbb{R}^n$, tel que $y = BB^T z$

on peut choisir $z = Bx$, alors $B^T z \neq 0$, et on a

$y = BB^T z = BB^T Bx = B(B^T B)x \in \mathrm{rang}(BB^T)$

② Pour montrer $\mathrm{range}(BB^T) \subseteq \mathrm{range}(B)$

$\forall\, w \in \mathrm{range}(BB^T)$, il existe $u \in \mathbb{R}^n$

$w = BB^T u$

on peut choisir $v = B^T u \in \mathbb{R}^m$, tel que

$w = Bv$

donc $w \in \mathrm{rang}(B)$

donc $\mathrm{range}(B) = \mathrm{range}(B^T B)$

日期:　　　/