Examen : Introduction à l'apprentissage automatique

18 novembre 2023

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. $\mathbb{1}$ le vecteur rempli de 1 (de taille adéquate).
2. $\mathbb{1}_A = \begin{cases} 1 & \text{si } A \text{ est vrai} \\ 0 & \text{sinon.} \end{cases}$
3. $\operatorname{card}(I)$ le cardinal de tout ensemble $I \subset \mathbb{N}$.
4. $\mathbf{I}_n$ la matrice identité de taille n (la taille peut varier).
5. $\operatorname{sign} : x \in \mathbb{R} \mapsto \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{sinon.} \end{cases}$
6. $\|A\|_F = \sqrt{\operatorname{tr}(AA^\top)}$ la norme de Frobénius de toute matrice A.

Exercice 1 (Algorithme EM, 3½ points)

Soient $(X_1, Y_1), \dots, (X_n, Y_n)$ des couples i.i.d. à valeurs dans $\mathbb{N} \times \{0, 1\}$ telles que

$$\begin{cases} Y_1 \sim \mathcal{B}(\pi^\star), & \pi^\star \in \,]0, 1[\\ X_1 \,\big|\, Y_1 \sim \mathcal{P}(\lambda^\star_{Y_1}), & \lambda^\star_{Y_1} > 0, \end{cases}$$

où $B(\pi^\star)$ est la loi de Bernoulli de paramètre $\pi^\star$ et $\mathcal{P}(\lambda)$ la loi de Poisson de paramètre $\lambda > 0$, de densité $x \in \mathbb{N} \mapsto \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda}$ par rapport à la mesure de comptage sur $\mathbb{N}$. Dans la suite, on souhaite partitionner $X_1, \dots, X_n$ via l'algorithme EM.

1. (1 point) On suppose d'abord observer $(X_1, Y_1), \dots, (X_n, Y_n)$. Définir, pour tout $\theta = (\pi, \lambda_1, \lambda_0) \in]0, 1[\times \mathbb{R}_+^* \times \mathbb{R}_+^*$ la log-vraisemblance $\ell(\theta, X_1, \dots, X_n, Y_1, \dots, Y_n)$ de θ au regard des observations et donner l'estimateur du maximum de vraisemblance de $\theta^\star = (\pi^\star, \lambda_1^\star, \lambda_0^\star)$.
2. (½ point) Déterminer la loi de $Y_1 \,\big|\, X_1$, notée $Q_{\theta^\star, X_1}$.

3. (1 point) On suppose maintenant n'observer que $X_1, \dots, X_n$ mais disposer d'un estimateur candidat $\hat\theta_0$. Soient alors $Z_1, \dots, Z_n$ telles que $Z_1 \,|\, \hat\theta_0, \dots, Z_n \,|\, \hat\theta_0$ sont i.i.d. et pour tout $i \in [\![1, n]\!]$, $Z_i \,|\, (X_1, \dots, X_n) = Z_i \,|\, (X_i, \hat\theta_0) \sim Q_{\hat\theta_0, X_i}$. Déterminer

$$\arg\max_{\theta \in]0,1[\times \mathbb{R}_+^* \times \mathbb{R}_+^*} F(\theta \,|\, \hat\theta_0) = \mathbb{E}\left[\ell(\theta, X_1, \dots, X_n, Z_1, \dots, Z_n \,\middle|\, (X_1, \dots, X_n)\right].$$

4. (1 point) Décrire l'algorithme EM produisant la suite d'estimateurs $(\hat\theta_t)_{t\geq 1}$ dans ce cas.

5. (1 point (bonus)) En raisonnant de manière générale et en appelant $m_{\theta^\star}$ la densité marginale de $\mathbf{X} = (X_1, \dots, X_n)$, montrer qu'à chaque iteration $t \geq 1$,

$$\log\left(m_{\hat\theta_{t+1}}(\mathbf{X})\right) - \log\left(m_{\hat\theta_t}(\mathbf{X})\right) \geq 0.$$

On devra faire intervenir un vecteur aléatoire $\mathbf{Z} = (Z_1, \dots, Z_n)$ tel que

$$\mathbf{Z} \,|\, \mathbf{X} \sim Q_{\hat\theta_t, X_1} \otimes \cdots \otimes Q_{\hat\theta_t, X_n}.$$

Exercice 2 (Clustering spectral, 3½ points)

Soient $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ un jeu de données, $s : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ une mesure de similarité symétrique et $W = (s(x_i, x_j))_{1\leq i,j\leq n}$ la matrice de similarité correspondante. Nous avons montré en cours que, pour n'importe quelle partition $(I_1, \dots, I_k)$ des indices $[\![1, n]\!]$

$$\mathrm{RatioCut}(I_1, \dots, I_k) = \mathrm{tr}(H^\top L H),$$

où $L = \left(\left[\sum_{\ell=1}^n W_{i,\ell}\right] \mathbb{1}_{i=j} - W_{i,j}\right)_{1\leq i,j\leq n}$ et $H = \left(\frac{\mathbb{1}_{i\in I_j}}{\sqrt{\mathrm{card}(I_j)}}\right)_{\substack{1\leq i\leq n \\ 1\leq j\leq k}}$.

1. (1 point) Rappeler l'algorithme de clustering spectral construit sur le RatioCut.

2. (1 point) Soit $J = (\mathbb{1}_{i\in I_j})_{\substack{1\leq i\leq n \\ 1\leq j\leq k}}$. Exprimer la matrice de projection orthogonale sur $\mathrm{range}(J)$, notée P_J, qui est l'unique matrice de taille $n \times n$ telle que pour tout $x \in \mathbb{R}^n$,

$$\{P_J x\} = \arg\min_{y\in\mathrm{range}(J)} \|x - y\|_{\ell_2}$$

et montrer que sa décomposition en éléments propres est définie par les vecteurs propres $u_j = \left(\frac{\mathbb{1}_{i\in I_j}}{\sqrt{\mathrm{card}(I_j)}}\right)_{1\leq i\leq n}$, $j = 1 \dots k$, de valeur propre commune 1.

3. (1½ points) Soit $\mathbb{S} = \left\{U \in \mathbb{R}^{n\times k} : U^\top U = \mathbf{I}_k\right\}$. On souhaite illustrer l'assertion « $\mathrm{range}(H)$ est l'espace vectoriel le plus proche de $\mathrm{range}(J)$ parmi ceux engendrés par les matrices de l'ensemble $\mathbb{S}$ ». Pour ce faire, montrer que

$$H \in \arg\min_{U\in\mathbb{S}} \|P_J - P_U\|_F,$$

où P_U est le projecteur orthogonale sur $\mathrm{range}(U)$.

Exercice 3 (Fonctions de perte pour la classification, 5½ points)

On considère un couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$ tel que $\eta : x \in \mathbb{R}^d \mapsto \mathbb{P}(Y = 1 \mid X = x) \in]0, 1[$. Soit maintenant $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe et différentiable autour de 0 telle que

$$\varphi'(0) < 0 \qquad \text{et} \qquad \arg\min_{u \in \mathbb{R}} \varphi(u) \neq \emptyset.$$

Remarquons qu'en particulier :

$$\forall u \in \mathbb{R} : \quad \varphi(u) \geq \varphi(0) + \varphi'(0)u.$$

1. (1½ points) En remarquant que $\varphi'(0) = \lim_{\substack{u \to 0 \\ u > 0}} \frac{\varphi(u) - \varphi(0)}{u}$, montrer que $\exists \bar{u} > 0$: $\varphi(0) > \varphi(\bar{u})$. En déduire que $\forall u \leq 0, \varphi(u) > \varphi(\bar{u})$, puis que
$$\arg\min_{u \in \mathbb{R}} \varphi(u) \subset \mathbb{R}_+^*.$$

2. (1½ points) Soient $\ell : \mathbb{R} \to \mathbb{R}$ une fonction de perte convexe, positive et différentiable autour de 0 telle que $\ell'(0) < 0$. Justifier que pour tout $x \in \mathbb{R}^d$
$$\varphi : u \in \mathbb{R} \mapsto \mathbb{E}\left[\ell(Yu) \mid X = x\right]$$
est coercive (*i.e.* $\lim_{-\infty} \varphi = \lim_{\infty} \varphi = \infty$). En appelant u_x^* un miniseur de φ sur $\mathbb{R}$, montrer que $f^* : x \mapsto u_x^*$ est minimiseur du risque
$$f \mapsto \mathbb{E}[\ell(Y f(X))]$$
et que $g^\star : x \mapsto \operatorname{sign}(f^*(x))$ est un classifieur de Bayes.

3. (1 point) On choisit $\ell : u \mapsto \max(0, 1 - u)^2$. Dessiner le graphe de φ puis exprimer $f^\star$ dans ce cas.

4. (1½ points) Proposer une manière d'estimer $g^\star$ par un classifieur linéaire fondé sur la minimisation d'un risque régularisé construit sur la perte ℓ et exprimer le gradient de ce risque.

5. (1 point (bonus)) Montrer que pour $\ell : u \mapsto \frac{1}{(1+\mathrm{e}^u)^2}$, $f^\star : x \mapsto \log\left(\frac{\eta(x)}{1-\eta(x)}\right)$. Conclure.

Exercice 4 (Classification à noyau, 7½ points)

Soient $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un jeu de données, $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ un noyau, $\mathcal{H}$ le RKHS associé et $\lambda > 0$. On s'intéresse à la construction d'un classifieur via la résolution du problème d'optimisation

$$\begin{array}{ll} \underset{h \in \mathcal{H}, \xi \in \mathbb{R}^n}{\text{minimiser}} & \dfrac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \dfrac{1}{2} \displaystyle\sum_{i=1}^n \xi_i^2 \\ \text{s.c.} & \forall i \in [\![1, n]\!] \begin{cases} Y_i h(X_i) \geq 1 - \xi_i & : \alpha_i \geq 0 \\ \xi_i \geq 0 & : \beta_i \geq 0. \end{cases} \end{array} \tag{P1}$$

dans le dual (on donne dans (P1) les multiplicateurs de Lagrange α_i et β_i associés à chaque contrainte).

1. (1 point) Les conditions de qualification de Slater sont-elles vérifiées ? Définir le lagrangien $\mathcal{L}$ associé à (P1) et expliciter, pour tous $\alpha \in \mathbb{R}_+^n$ et $\beta \in \mathbb{R}_+^n$, les conditions de stationarité primale en $(h, \xi) \in \mathcal{H} \times \mathbb{R}^n$:
$$\nabla_h \mathcal{L}(h, \xi, \alpha, \beta) = 0 \qquad \text{et} \qquad \nabla_\xi \mathcal{L}(h, \xi, \alpha, \beta) = 0.$$
2. (1½ points) Montrer qu'un problème dual à (P1) est
$$\underset{\alpha \in \mathbb{R}_+^n,\, \beta \in \mathbb{R}_+^n}{\text{maximiser}} \; -\frac{1}{2\lambda}\alpha^\top Q \alpha + \mathbb{1}^\top \alpha - \frac{1}{2}\left\|\alpha + \beta\right\|_{\ell_2}^2 ,$$
où Q est une matrice à préciser, puis que ce problème est équivalent (au sens où connaissant les solutions de l'un, on peut déterminer celles de l'autre et vice versa) à
$$\underset{\alpha \in \mathbb{R}_+^n}{\text{minimiser}} \; \frac{1}{2\lambda}\alpha^\top P \alpha - \mathbb{1}^\top \alpha, \tag{P2}$$
où P est une matrice à préciser.
3. (1 point) Montrer que P est symétrique et semi-définie positive. Que peut-on en déduire de (P2) ?
4. (1 point) Expliciter les étapes d'un algorithme de résolution de (P2) de type « descente par coordonnée ».
5. (1½ points) Énoncer les conditions KKT pour des candidats solutions $(h^\star, \xi^\star)$ et $(\alpha^\star, \beta^\star)$ et en déduire une classifieur issu de la résolution de (P1).
6. (1½ points) Justifier que, connaissant $h^\star$, on peut choisir $\xi_i^\star = \max\left(0, 1 - Y_i h^\star(x_i)\right)$ pour tout $i \in [\![1, n]\!]$. En déduire que pour tout $i \in [\![1, n]\!]$, si $Y_i h^\star(x_i) > 1$, alors $\alpha_i^\star = 0$.
7. (1 point (bonus)) Proposer un critère d'arrêt pour l'algorithme itératif de la question 4. Le justifier rapidement et expliciter le calcul.

Exercice 1 (Algorithme EM, 3½ points)

Soient $(X_1, Y_1), \dots, (X_n, Y_n)$ des couples i.i.d. à valeurs dans $\mathbb{N} \times \{0,1\}$ telles que

$$\begin{cases} Y_1 \sim \mathcal{B}(\pi^\star), & \pi^\star \in]0,1[\\ X_1 \,|\, Y_1 \sim \mathcal{P}(\lambda^\star_{Y_1}), & \lambda^\star_{Y_1} > 0, \end{cases}$$

où $B(\pi^\star)$ est la loi de Bernoulli de paramètre $\pi^\star$ et $\mathcal{P}(\lambda)$ la loi de Poisson de paramètre $\lambda > 0$, de densité $x \in \mathbb{N} \mapsto \frac{\lambda^x}{x!} e^{-\lambda}$ par rapport à la mesure de comptage sur $\mathbb{N}$. Dans la suite, on souhaite partitionner $X_1, \dots, X_n$ via l'algorithme EM.

1. (1 point) On suppose d'abord observer $(X_1, Y_1), \dots, (X_n, Y_n)$. Définir, pour tout $\theta = (\pi, \lambda_1, \lambda_0) \in]0,1[\times \mathbb{R}^*_+ \times \mathbb{R}^*_+$ la log-vraisemblance $\ell(\theta, X_1, \dots, X_n, Y_1, \dots, Y_n)$ de θ au regard des observations et donner l'estimateur du maximum de vraisemblance de $\theta^\star = (\pi^\star, \lambda^\star_1, \lambda^\star_0)$.

(1) $\forall i \in [\![1,n]\!]$, $\mathbb{P}(Y_i = 1) = \pi$ $\quad \mathbb{P}(Y_i = 0) = 1-\pi$

$\mathbb{P}(X_i = x \,|\, Y_i = 1) = \frac{\lambda_1^x}{x!} \cdot e^{-\lambda_1}$ $\quad \mathbb{P}(X_i = x \,|\, Y_i = 0) = \frac{\lambda_0^x}{x!} \cdot e^{-\lambda_0}$

donc

$$f_{(X,Y)}(x,y) = \left(\pi \cdot \frac{\lambda_1^x}{x!} \cdot e^{-\lambda_1}\right)^{\mathbb{1}_{\{y=1\}}} \cdot \left((1-\pi) \cdot \frac{\lambda_0^x}{x!} \cdot e^{-\lambda_0}\right)^{\mathbb{1}_{\{y=0\}}}$$

$$\log f_{(X,Y)}(x,y) = \mathbb{1}_{\{y=1\}} \cdot (\log \pi + x \cdot \log \lambda_1 - \log x! - \lambda_1)$$
$$+ \mathbb{1}_{\{y=0\}} \cdot (\log(1-\pi) + x \cdot \log \lambda_0 - \log x! - \lambda_0)$$

donc

$$\ell(\theta, X_1, \dots, X_n, Y_1, \dots, Y_n) = \sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}} \cdot (\log \pi + X_i \cdot \log \lambda_1 - \log X_i! - \lambda_1)$$
$$+ \sum_{i=1}^n \mathbb{1}_{\{Y_i=0\}} \cdot (\log(1-\pi) + X_i \cdot \log \lambda_0 - \log X_i! - \lambda_0)$$

$$\frac{\partial \ell}{\partial \pi} = 0 \implies \sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}} \cdot \frac{1}{\pi} - \sum_{i=1}^n \mathbb{1}_{\{Y_i=0\}} \frac{1}{1-\pi} = 0$$

$$\pi^* = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}}$$

$$\frac{\partial \ell}{\partial \lambda_1} = 0 \implies \sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}} \cdot \left(\frac{X_i}{\lambda_1} - 1\right) = 0$$

$$\lambda_1^* = \frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}} \cdot X_i}{\sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}}}$$

$$\frac{\partial \ell}{\partial \lambda_0} = 0 \implies \lambda_0^* = \frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i=0\}} \cdot X_i}{\sum_{i=1}^n \mathbb{1}_{\{Y_i=0\}}}$$

2. (½ point) Déterminer la loi de $Y_1 \mid X_1$, notée $Q_{\theta^\star, X_1}$.

(2) $$\mathbb{P}(Y_1=1 \mid X_1) = \frac{\mathbb{P}(Y_1=1, X_1=x)}{\mathbb{P}(X_1=x)} = \frac{\pi \cdot \lambda_1^{X_1} \cdot e^{-\lambda_1}}{\pi \cdot \lambda_1^{X_1} \cdot e^{-\lambda_1} + (1-\pi)\cdot \lambda_0^{X_1} \cdot e^{-\lambda_0}} = q_{\theta, X_1}$$

$$Y_1 \mid X_1 \sim \mathcal{B}\left(\frac{\pi \cdot \lambda_1^{X_1} \cdot e^{-\lambda_1}}{\pi \cdot \lambda_1^{X_1} \cdot e^{-\lambda_1} + (1-\pi)\cdot \lambda_0^{X_1} \cdot e^{-\lambda_0}} \right)$$

3. (1 point) On suppose maintenant n'observer que $X_1, \ldots, X_n$ mais disposer d'un estimateur candidat $\hat{\theta}_0$. Soient alors $Z_1, \ldots, Z_n$ telles que $Z_1 \mid \hat{\theta}_0, \ldots, Z_n \mid \hat{\theta}_0$ sont i.i.d. et pour tout $i \in [\![1, n]\!]$, $Z_i \mid (X_1, \ldots, X_n) = Z_i \mid (X_i, \hat{\theta}_0) \sim Q_{\hat{\theta}_0, X_i}$. Déterminer

$$\arg\max_{\theta \in]0,1[\times \mathbb{R}_+^* \times \mathbb{R}_+^*} F(\theta \mid \hat{\theta}_0) = \mathbb{E}\left[\ell(\theta, X_1, \ldots, X_n, Z_1, \ldots, Z_n \mid (X_1, \ldots, X_n) \right].$$

(3) $$\begin{aligned} F(\theta \mid \hat{\theta}_0) &= \mathbb{E}\Big[\sum_{i=1}^n \mathbb{1}_{\{Z_i=1\}} \cdot (\log \pi + X_i \cdot \log \lambda_1 - \log X_i! - \lambda_1) \\ &\quad + \sum_{i=1}^n \mathbb{1}_{\{Z_i=0\}} \cdot (\log(1-\pi) + X_i \cdot \log \lambda_0 - \log X_i! - \lambda_0) \,\Big|\, (X_1, \ldots, X_n) \Big] \\ &= \sum_{i=1}^n q_{\hat{\theta}_0, X_i} \cdot (\log \pi + X_i \cdot \log \lambda_1 - \log X_i! - \lambda_1) \\ &\quad + \sum_{i=1}^n (1 - q_{\hat{\theta}_0, X_i}) \cdot (\log(1-\pi) + X_i \cdot \log \lambda_0 - \log X_i! - \lambda_0) \end{aligned}$$

donc

$$\frac{\partial F}{\partial \pi} = 0 \implies \pi^* = \frac{1}{n} \sum_{i=1}^n \frac{\hat{\pi}_0 \cdot \hat{\lambda}_1^{X_i} \cdot e^{-\hat{\lambda}_1}}{\hat{\pi}_0 \cdot \hat{\lambda}_1^{X_i} \cdot e^{-\hat{\lambda}_1} + (1-\hat{\pi}_0) \cdot \hat{\lambda}_0^{X_i} \cdot e^{-\hat{\lambda}_0}}$$

$$\frac{\partial F}{\partial \lambda_1} = 0 \implies \lambda_1^* = \frac{\sum_{i=1}^n q_{\hat{\theta}_0, X_i} \cdot X_i}{\sum_{i=1}^n q_{\hat{\theta}_0, X_i}}$$

$$\frac{\partial F}{\partial \lambda_0} = 0 \implies \lambda_0^* = \frac{\sum_{i=1}^n (1 - q_{\hat{\theta}_0, X_i}) \cdot X_i}{\sum_{i=1}^n (1 - q_{\hat{\theta}_0, X_i})}$$

4. (1 point) Décrire l'algorithme EM produisant la suite d'estimateurs $(\hat{\theta}_t)_{t\geq 1}$ dans ce cas.

(4) L'algorithme EM

Input : $X_1, \dots, X_n$

Initialiser $\hat{\theta}_0 = (\hat{\pi}_0, \hat{\lambda}_1^{(0)}, \hat{\lambda}_0^{(0)})$

pour le $i^{\text{ième}}$ étape :

$$\hat{\pi}_i \longleftarrow \frac{1}{n}\sum_{i=1}^{n} q_{\hat{\theta}_{i-1}, X_i}$$

$$\hat{\lambda}_1^{(i)} \longleftarrow \frac{\sum_{i=1}^{n} q_{\hat{\theta}_{i-1}, X_i} \cdot X_i}{\sum_{i=1}^{n} q_{\hat{\theta}_{i-1}, X_i}}$$

$$\hat{\lambda}_0^{(i)} \longleftarrow \frac{\sum_{i=1}^{n} (1 - q_{\hat{\theta}_{i-1}, X_i}) \cdot X_i}{\sum_{i=1}^{n} (1 - q_{\hat{\theta}_{i-1}, X_i})}$$

Output $q_{\hat{\theta}_i, X_1}, \dots, q_{\hat{\theta}_i, X_n}$

5. (1 point (bonus)) En raisonnant de manière générale et en appelant $m_{\theta^\star}$ la densité marginale de $\mathbf{X} = (X_1, \ldots, X_n)$, montrer qu'à chaque iteration $t \geq 1$,

$$\log\left(m_{\hat{\theta}_{t+1}}(\mathbf{X})\right) - \log\left(m_{\hat{\theta}_t}(\mathbf{X})\right) \geq 0.$$

On devra faire intervenir un vecteur aléatoire $\mathbf{Z} = (Z_1, \ldots, Z_n)$ tel que

$$\mathbf{Z} \,|\, \mathbf{X} \sim Q_{\hat{\theta}_t, X_1} \otimes \cdots \otimes Q_{\hat{\theta}_t, X_n}.$$

(5) $\mathbb{P}(X=x) = \pi \cdot \frac{\lambda_1^x}{x!} \cdot e^{-\lambda_1} + (1-\pi) \cdot \frac{\lambda_0^x}{x!} \cdot e^{-\lambda_0}$

Donc

$$m_{\hat{\theta}_t}(X) = \prod_{i=1}^{n} \left(q_{\hat{\theta}_t, X_i} \cdot \frac{\lambda_1^{X_i}}{X_i!} \cdot e^{-\lambda_1} + (1 - q_{\hat{\theta}_t, X_i}) \cdot \frac{\lambda_0^{X_i}}{X_i!} \cdot e^{-\lambda_0} \right)$$

$$\log(m_{\hat{\theta}_t}(X)) = \sum_{i=1}^{n} \Big(\log q_{\hat{\theta}_t, X_i} + X_i \cdot \log \hat{\lambda}_1^{(t)} - \log(X_i!) - \hat{\lambda}_1^{(t)}$$
$$+ \log(1 - q_{\hat{\theta}_t, X_i}) + X_i \cdot \log \hat{\lambda}_0^{(t)} - \log(X_i!) - \hat{\lambda}_0^{(t)} \Big)$$

$$\log(m_{\hat{\theta}_{t+1}}(X)) - \log(m_{\hat{\theta}_t}(X)) = \sum_{i=1}^{n} \Big(\log \frac{q_{\hat{\theta}_{t+1}, X_i}}{q_{\hat{\theta}_t, X_i}} + X_i \cdot \log \frac{\hat{\lambda}_1^{(t+1)}}{\hat{\lambda}_1^{(t)}} - (\hat{\lambda}_1^{(t+1)} - \hat{\lambda}_1^{(t)})$$
$$+ \log \frac{1 - q_{\hat{\theta}_{t+1}, X_i}}{1 - q_{\hat{\theta}_t, X_i}} + X_i \cdot \log \frac{\hat{\lambda}_0^{(t+1)}}{\hat{\lambda}_0^{(t+1)}} - (\hat{\lambda}_0^{(t+1)} - \hat{\lambda}_0^{(t)}) \Big)$$

$$\ldots$$

$$\geqslant 0$$

Exercice 2 (Clustering spectral, 3½ points)

Soient $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ un jeu de données, $s : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ une mesure de similarité symétrique et $W = (s(x_i, x_j))_{1 \le i,j \le n}$ la matrice de similarité correspondante. Nous avons montré en cours que, pour n'importe quelle partition $(I_1, \ldots, I_k)$ des indices $[\![1, n]\!]$

$$\mathrm{RatioCut}(I_1, \ldots, I_k) = \mathrm{tr}(H^\top L H),$$

où $L = ([\sum_{\ell=1}^n W_{i,\ell}] \mathbb{1}_{i=j} - W_{i,j})_{1 \le i,j \le n}$ et $H = \left(\frac{\mathbb{1}_{i \in I_j}}{\sqrt{\mathrm{card}(I_j)}} \right)_{\substack{1 \le i \le n \\ 1 \le j \le k}}$.

1. (1 point) Rappeler l'algorithme de clustering spectral construit sur le RatioCut.

(1) L'algorithme de clustering spectral (RatioCut)

Unnormalized Spectral Clustering

Input : $W \in \mathbb{R}^{n \times n}$ (adjency matrix)

$L \longleftarrow$ Laplacian de W

$H \longleftarrow$ k minor eigenvectors of L as columns

$Y_i \longleftarrow$ i^{th} row of H

$(\hat{C}_1, \ldots, \hat{C}_k) \longleftarrow$ output of k-means algorithm based on $(Y_1, \ldots, Y_n)$

Output : $(\hat{C}_1, \ldots, \hat{C}_k)$

2. (1 point) Soit $J = (\mathbb{1}_{i \in I_j})_{\substack{1 \le i \le n \\ 1 \le j \le k}}$. Exprimer la matrice de projection orthogonale sur range(J), notée P_J, qui est l'unique matrice de taille $n \times n$ telle que pour tout $x \in \mathbb{R}^n$,

$$\{P_J x\} = \arg\min_{y \in \text{range}(J)} \|x - y\|_{\ell_2}$$

et montrer que sa décomposition en éléments propres est définie par les vecteurs propres $u_j = \left(\frac{\mathbb{1}_{i \in I_j}}{\sqrt{\text{card}(I_j)}}\right)_{1 \le i \le n}$, $j = 1 \dots k$, de valeur propre commune 1.

(2)

$$J = \begin{bmatrix} \mathbb{1}_{1 \in I_1} & \mathbb{1}_{1 \in I_2} & \cdots & \mathbb{1}_{1 \in I_k} \\ \mathbb{1}_{2 \in I_1} & \mathbb{1}_{2 \in I_2} & \cdots & \mathbb{1}_{2 \in I_k} \\ \vdots & \vdots & & \vdots \\ \mathbb{1}_{n \in I_1} & \mathbb{1}_{n \in I_2} & \cdots & \mathbb{1}_{n \in I_k} \end{bmatrix}_{n \times k} \xRightarrow{\text{变换}} J = \Sigma \cdot \begin{bmatrix} 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & 1 & \\ & & \vdots & \\ & & 1 & \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{bmatrix}_{n \times k}$$

on a

$$P_J = J(J^TJ)^{-1}J^T$$

$$= J \cdot \begin{bmatrix} \text{card}(I_1) & & \\ & \ddots & \\ & & \text{card}(I_k) \end{bmatrix}_{k \times k}^{-1} J^T$$

$$= \Sigma \cdot \begin{bmatrix} \frac{1}{\text{card}(I_1)} & \cdots & \frac{1}{\text{card}(I_1)} & & & \\ \vdots & & \vdots & & & \\ \frac{1}{\text{card}(I_1)} & \cdots & \frac{1}{\text{card}(I_1)} & & & \\ & & & \ddots & & \\ & & & & \frac{1}{\text{card}(I_k)} & \cdots & \frac{1}{\text{card}(I_k)} \\ & & & & \vdots & & \vdots \\ & & & & \frac{1}{\text{card}(I_k)} & \cdots & \frac{1}{\text{card}(I_k)} \end{bmatrix} \Sigma^T$$

Donc $\forall j \in [\![1, k]\!]$

$$P_J \cdot u_j = u_j$$

3. (1½ points) Soit $\mathbb{S} = \{U \in \mathbb{R}^{n\times k} : U^\top U = \mathbf{I}_k\}$. On souhaite illustrer l'assertion « range(H) est l'espace vectoriel le plus proche de range(J) parmi ceux engendrés par les matrices de l'ensemble $\mathbb{S}$ ». Pour ce faire, montrer que

$$H \in \arg\min_{U\in\mathbb{S}} \|P_J - P_U\|_F \,,$$

où P_U est le projecteur orthogonale sur range(U).

(3) 0 $\quad \underset{U\in\$}{\operatorname{argmin}} \|P_J - P_U\|_F = \underset{\substack{U\in\mathbb{R}^{n\times k} \\ U^T U = I_k}}{\operatorname{argmin}} \sqrt{\operatorname{tr}((P_J-P_U)(P_J-P_U)^T)}$

alors on a $\|P_J - P_U\|_F \geqslant 0$

comme $P_H = P_J$, alors on a

$\|P_J - P_H\|_F = 0$

donc $H \in \underset{U\in\$}{\operatorname{argmin}} \|P_J - P_U\|_F$

Exercice 3 (Fonctions de perte pour la classification, 5½ points)

On considère un couple aléatoire (X, Y) à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$ tel que $\eta : x \in \mathbb{R}^d \mapsto \mathbb{P}(Y = 1 \mid X = x) \in]0,1[$. Soit maintenant $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe et différentiable autour de 0 telle que

$$\varphi'(0) < 0 \qquad \text{et} \qquad \arg\min_{u \in \mathbb{R}} \varphi(u) \neq \emptyset.$$

Remarquons qu'en particulier :

$$\forall u \in \mathbb{R} : \quad \varphi(u) \geq \varphi(0) + \varphi'(0)u.$$

1. (1½ points) En remarquant que $\varphi'(0) = \lim_{\substack{u \to 0 \\ u > 0}} \frac{\varphi(u) - \varphi(0)}{u}$, montrer que $\exists \bar{u} > 0$: $\varphi(0) > \varphi(\bar{u})$. En déduire que $\forall u \leq 0, \varphi(u) > \varphi(\bar{u})$, puis que
$$\arg\min_{u \in \mathbb{R}} \varphi(u) \subset \mathbb{R}_+^*.$$

---

1) Par l'hypothése, $\underset{u \in \mathbb{R}}{\text{argmin}}\, \varphi(u) \neq \phi$

donc il existe $u \in \mathbb{R}$ tel que $\varphi'(u) = 0$

et comme φ est convexe et $\varphi'(0) < 0$

alors il existe $\bar{u} > 0$, tel que $\varphi'(\bar{u}) = 0$

et $\varphi(0) > \varphi(\bar{u})$

Et comme $\varphi'(0) < 0$, φ est convexe

alors $\forall u \leq 0$, $\varphi'(u) < 0$

donc φ est strictement décroissant dans $]-\infty, 0]$

donc $\forall u \leq 0$, $\varphi(u) \geqslant \varphi(0) > \varphi(\bar{u})$

Donc $\underset{u \in \mathbb{R}}{\text{argmin}}\, \varphi(u) \subset \mathbb{R}^+$

2. ($1\frac{1}{2}$ points) Soient $\ell : \mathbb{R} \to \mathbb{R}$ une fonction de perte convexe, positive et différentiable autour de 0 telle que $\ell'(0) < 0$. Justifier que pour tout $x \in \mathbb{R}^d$

$$\varphi : u \in \mathbb{R} \mapsto \mathbb{E}\left[\ell(Yu) \mid X = x\right]$$

est coercive (*i.e.* $\lim_{-\infty} \varphi = \lim_{\infty} \varphi = \infty$). En appelant u_x^* un miniseur de φ sur $\mathbb{R}$, montrer que $f^* : x \mapsto u_x^*$ est minimiseur du risque

$$f \mapsto \mathbb{E}[\ell(Yf(X))]$$

et que $g^\star : x \mapsto \mathrm{sign}(f^*(x))$ est un classifieur de Bayes.

(2) $\lim\limits_{u\to+\infty} \varphi(u) = \lim\limits_{u\to+\infty} \mathbb{E}\left[\ell(Yu) \mid X=x\right]$

$= \mathbb{E}\left[\lim\limits_{u\to+\infty} \ell(Yu) \mid X=x\right]$

comme ℓ est convexe, positive et différentiable

alors $\lim\limits_{u\to+\infty} \ell(u) = \lim\limits_{u\to-\infty} \ell(u) = +\infty$

donc $\lim\limits_{u\to+\infty} \ell(Yu) = \lim\limits_{u\to-\infty} \ell(Yu) = +\infty$, $\forall Y \in \{\pm 1\}$

donc $\lim\limits_{u\to+\infty} \varphi(u) = \lim\limits_{u\to-\infty} \varphi(u) = +\infty$

c-à-d φ est coercive

Comme u_x^* est un miniseur de φ, alors

$\mathcal{R}(f^*) = \mathbb{E}\left[\ell(Yf^*(X))\right]$

$= \mathbb{E}\left[\ell(Yf^*(x)) \mid X=x\right]$

$= \mathbb{E}\left[\ell(Y \cdot u_x^*) \mid X=x\right]$

$= \varphi(u_x^*)$

$= \min\limits_{u\in\mathbb{R}} \varphi(u) = \min\limits_{f} \mathcal{R}(f)$

donc f^* est un minimiseur du risque

Si $Y=1$, $\varphi(u) = \mathbb{E}[\ell(u) | X=x]$

d'après (1) $u_x^* \in \mathbb{R}_+$ alors $g^*(x) = 1$

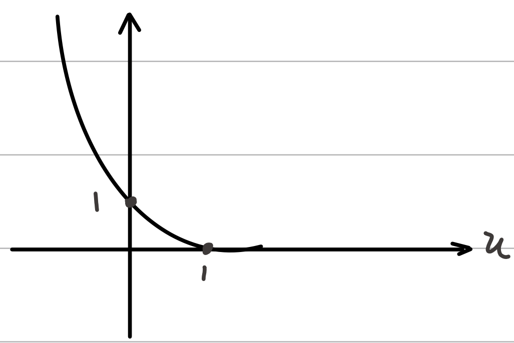
Si $Y=-1$, $\varphi(u) = \mathbb{E}[\ell(-u) | X=x]$

d'après (1) $u_x^* \in \mathbb{R}_-$ alors $g^*(x) = -1$

Donc g^* est un classifieur de Bayes.

3. (1 point) On choisit $\ell : u \mapsto \max(0, 1-u)^2$. Dessiner le graphe de φ puis exprimer $f^\star$ dans ce cas.

(3) Le graphe



$$R(f) = \mathbb{E}[\ell(Yf(X))] = \mathbb{E}[\max(0, 1-Yf(X))^2]$$

comme f^* est minimiseur de $R(f)$

alors $Yf^*(X) \geqslant 1$

c-à-d si $Y=1$, $f^*(X) \geqslant 1$

si $Y=-1$, $f^*(X) \leqslant -1$

估计 $f^*(x) = w^T x + b$

4. (1½ points) Proposer une manière d'estimer $g^\star$ par un classifieur linéaire fondé sur la minimisation d'un risque régularisé construit sur la perte ℓ et exprimer le gradient de ce risque.

(4) Pour estimer $g^*(x) = \text{sign}(f^*(x)) = \text{sign}(w^{*T}x + b^*)$

on doit estimer (w^*, b^*)

① On prend N échantillons $(X_1, Y_1), \dots, (X_n, Y_n)$

qui sont i.i.d. de (X, Y)

② les paramétres (w^*, b^*) est solution de

$$\underset{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}{\text{minimise}} \ \frac{\lambda}{2}\|w\|_{\ell_2}^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - Y_i f(X_i))^2$$

$$\iff \underset{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}}{\text{minimise}} \ \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{\lambda n}\sum_{i=1}^{n} \xi_i^2$$

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

D'après (3), pour f^*, on a $Y_i f^*(X_i) \geq 1$, $\forall i \in [\![1, n]\!]$

c-à-d $\xi_i = 0$, $\forall i \in [\![1, n]\!]$

Donc on a le nouveau probléme

$$\underset{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}{\text{minimise}} \ \frac{1}{2}\|w\|_{\ell_2}^2$$

$$\text{s.c. } \forall i \in [\![1, n]\!], \ Y_i(w^T X_i + b) \geq 1$$

Alors on a $\hat{g}(x) = \text{sign}(\hat{w}^T x + \hat{b})$

日期:　　　/

Pour $R(f) = E\left[\max(0, 1 - Y f(X))^2\right]$

on a

$$R_n(w,b) = \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - Y_i(w^T X_i + b))^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} (1 - Y_i(w^T X_i + b))^2 \cdot \mathbb{1}\{Y_i(w^T X_i + b) \leq 1\}$$

$$\nabla_w R_n(w,b) = -\frac{2}{n}\sum_{i=1}^{n} Y_i X_i \cdot (1 - Y_i(w^T X_i + b)) \cdot \mathbb{1}\{Y_i(w^T X_i + b) \leq 1\}$$
$$\nabla_b R_n(w,b) = -\frac{2}{n}\sum_{i=1}^{n} Y_i \cdot (1 - Y_i(w^T X_i + b)) \cdot \mathbb{1}\{Y_i(w^T X_i + b) \leq 1\}$$

5. (1 point (bonus)) Montrer que pour $\ell : u \mapsto \frac{1}{(1+e^u)^2}$, $f^\star : x \mapsto \log\left(\frac{\eta(x)}{1-\eta(x)}\right)$. Conclure.

(5) Pour $\ell(u) = \frac{1}{(1+e^u)^2}$ $\quad \ell'(u) = -\frac{2\cdot e^u}{(1+e^u)^3}$

on a

$$R(f) = \mathbb{E}[\ell(Yf(X))]$$
$$= \mathbb{E}[\eta(X)\cdot\ell(f(X)) + (1-\eta(X))\cdot\ell(-f(X))]$$

pour fonction

$$C_\eta(u) = \eta\cdot\ell(u) + (1-\eta)\cdot\ell(-u)$$

on a

$$C_\eta'(u) = \eta\cdot\ell'(u) + (1-\eta)\cdot\ell'(-u)$$
$$= -\eta\cdot\frac{2\cdot e^u}{(1+e^u)^3} + (1-\eta)\cdot\frac{2\cdot e^{-u}}{(1+e^{-u})^3}$$

alors

$$C_\eta'(u) = 0 \implies \eta\cdot\frac{2\cdot e^u}{(1+e^u)^3} = (1-\eta)\cdot\frac{2\cdot e^{-u}}{(1+e^{-u})^3}$$
$$\frac{\eta\cdot e^u}{(1+e^u)^3} = \frac{(1-\eta)\cdot e^{2u}}{(1+e^u)^3}$$
$$e^u = \frac{\eta}{1-\eta}$$
$$u^* = \log\left(\frac{\eta}{1-\eta}\right)$$

donc

$$f^*(x) = \log\left(\frac{\eta(x)}{1-\eta(x)}\right)$$

Exercice 4 (Classification à noyau, 7½ points)

Soient $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un jeu de données, $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ un noyau, $\mathcal{H}$ le RKHS associé et $\lambda > 0$. On s'intéresse à la construction d'un classifieur via la résolution du problème d'optimisation

$$\begin{aligned} &\underset{h \in \mathcal{H},\, \xi \in \mathbb{R}^n}{\text{minimiser}} \ \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \frac{1}{2}\sum_{i=1}^n \xi_i^2 \\ &\text{s.c.} \quad \forall i \in [\![1, n]\!] \begin{cases} Y_i h(X_i) \geq 1 - \xi_i & : \alpha_i \geq 0 \\ \xi_i \geq 0 & : \beta_i \geq 0. \end{cases} \end{aligned} \tag{P1}$$

dans le dual (on donne dans (P1) les multiplicateurs de Lagrange α_i et β_i associés à chaque contrainte).

1. (1 point) Les conditions de qualification de Slater sont-elles vérifiées ? Définir le lagrangien $\mathcal{L}$ associé à (P1) et expliciter, pour tous $\alpha \in \mathbb{R}_+^n$ et $\beta \in \mathbb{R}_+^n$, les conditions de stationarité primale en $(h, \xi) \in \mathcal{H} \times \mathbb{R}^n$:

$$\nabla_h \mathcal{L}(h, \xi, \alpha, \beta) = 0 \qquad \text{et} \qquad \nabla_\xi \mathcal{L}(h, \xi, \alpha, \beta) = 0.$$

(1) Les conditions de qualification de Slater

$\forall i \in [\![1, n]\!]$, $g_i(h, \xi) = 1 - \xi_i - Y_i h(X_i)$

$m_i(\xi) = -\xi_i$

alors $\exists\, \tilde{\xi} = 2\mathbb{1}$, $\tilde{h}(x) = 0$, tel que

$\forall i \in [\![1, n]\!]$, $g_i(\tilde{h}, \tilde{\xi}) < 0$

$m_i(\tilde{\xi}) < 0$

donc les conditions de qualification de Slater sont vérifiés.

Lagrangien $\mathcal{L}$

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

$$\nabla_h \mathcal{L}(h, \xi, \alpha, \beta) = \lambda \cdot h - \sum_{i=1}^n \alpha_i \cdot Y_i \cdot k(\cdot, X_i)$$

$$\nabla_h \mathcal{L}(h, \xi, \alpha, \beta) = 0 \implies \hat{h} = \frac{1}{\lambda}\sum_{i=1}^n \alpha_i \cdot Y_i \cdot k(\cdot, X_i)$$

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

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

2. (1½ points) Montrer qu'un problème dual à (P1) est

$$\underset{\alpha\in\mathbb{R}^n_+,\,\beta\in\mathbb{R}^n_+}{\text{maximiser}} \ -\frac{1}{2\lambda}\alpha^\top Q\alpha + \mathbb{1}^\top\alpha - \frac{1}{2}\|\alpha+\beta\|^2_{\ell_2},$$

où Q est une matrice à préciser, puis que ce problème est équivalent (au sens où connaissant les solutions de l'un, on peut déterminer celles de l'autre et vice versa) à

$$\underset{\alpha\in\mathbb{R}^n_+}{\text{minimiser}} \ \frac{1}{2\lambda}\alpha^\top P\alpha - \mathbb{1}^\top\alpha, \tag{P2}$$

où P est une matrice à préciser.

---

(2) Par définition

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

$$= \frac{\lambda}{2}\cdot\Big\|\frac{1}{\lambda}\sum_{i=1}^n \alpha_i\cdot Y_i\cdot k(\cdot,X_i)\Big\|^2_{\mathcal{H}} + \frac{1}{2}(\alpha+\beta)^T(\alpha+\beta)$$

$$+\sum_{i=1}^n \alpha_i\cdot\Big(1-\alpha_i-\beta_i - Y_i\cdot\frac{1}{\lambda}\sum_{j=1}^n \alpha_j\cdot Y_j\cdot k(X_i,X_j)\Big) - \beta^T(\alpha+\beta)$$

$$= \frac{1}{2\lambda}\cdot\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j Y_i Y_j\cdot k(X_i,X_j) - \frac{1}{2}\|\alpha+\beta\|^2_{\ell_2} + \sum_{i=1}^n \alpha_i$$

$$-\frac{1}{\lambda}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j Y_i Y_j\cdot k(X_i,X_j)$$

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

$$= -\frac{1}{2\lambda}\alpha^T Q\alpha - \frac{1}{2}\|\alpha+\beta\|^2_{\ell_2} + \mathbb{1}^T\alpha$$

Le problème dual à (P_1) est

$$\underset{\alpha\in\mathbb{R}^n_+,\,\beta\in\mathbb{R}^n_+}{\text{maximiser}} \ -\frac{1}{2\lambda}\alpha^T Q\alpha - \frac{1}{2}\|\alpha+\beta\|^2_{\ell_2} + \mathbb{1}^T\alpha$$

Pour $\varphi(\beta) = -\frac{1}{2}\|\alpha+\beta\|^2_{\ell_2}$, $\nabla_\beta\varphi = -\beta^T(\alpha+\beta)$

$$\nabla_\beta\varphi = 0 \implies \beta^* = -\alpha \leq 0$$

comme $\varphi(\beta)$ est concave, on a

$$\max \varphi(u) = \varphi(0) = -\frac{1}{2}\|\alpha\|_{\ell_2}^2$$

Donc le probléme dual est équivalent à

$$\underset{\alpha \in \mathbb{R}_+^n}{\text{maximiser}} \ -\frac{1}{2\lambda}\alpha^T Q \alpha + \mathbb{1}^T \alpha - \frac{1}{2}\alpha^T\alpha$$

$$\Longleftrightarrow$$

$$\underset{\alpha \in \mathbb{R}_+^n}{\text{minimiser}} \ \frac{1}{2\lambda}\alpha^T (Q + \lambda \cdot I_n)\alpha - \mathbb{1}^T\alpha$$

avec $P = Q + \lambda \cdot I_n$

3. (1 point) Montrer que P est symétrique et semi-définie positive. Que peut-on en déduire de (P2) ?

(3) D'après (2), $P = Q + \lambda I_n$ avec $Q_{ij} = Y_i Y_j \cdot k(X_i, X_j)$

Symétrique

comme λI_n est symétrique, Q est symétrique

on a $P = Q + \lambda I_n$ est symétrique

Semi-définie positive

comme $\lambda > 0$, on a λI_n est semi-définie positive

d'après la définition, $K = (k(X_i, X_j))_{\substack{1 \le i \le n \\ 1 \le j \le n}}$ est S.D.P

- et comme $\forall i, j$, $Y_i Y_j = Y_j Y_i = \pm 1$

on a Q est semi-définie positive

donc $P = Q + \lambda I_n$ est semi-définie positive

Pour (P_2), comme P est semi-définie positive,
alors $\frac{1}{2}\alpha^T P\alpha - \mathbb{1}^T\alpha$ est convexe
donc il existe un minimiseur pour $\frac{1}{2}\alpha^T P\alpha - \mathbb{1}^T\alpha$

$$\nabla_\alpha\left(\frac{1}{2}\alpha^T P\alpha - \mathbb{1}^T\alpha\right) = P\alpha - \mathbb{1}$$

on a $\alpha^* = P^{-1}\cdot\mathbb{1}$ est une solution de (P_2)

4. (1 point) Expliciter les étapes d'un algorithme de résolution de (P2) de type « descente par coordonnée ».

(4) <u>Algorithm</u> : Sequential minimal optimization

Input : $C > 0$, $k : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$, $\{(X_i, Y_i)\}_{1\le i\le n}$

$Q \longleftarrow (Y_i Y_j \cdot k(X_i, X_j))_{1\le i,j\le n}$

while not converge do :

find α_i for which KKT condition are violated

pick $\alpha_j \neq \alpha_i$ at random

solve (P_2) with respect to (α_i, α_j) with all other fixed

end while

Output : $(\alpha_1, \dots, \alpha_n)$

5. ($1\frac{1}{2}$ points) Énoncer les conditions KKT pour des candidats solutions $(h^\star, \xi^\star)$ et $(\alpha^\star, \beta^\star)$ et en déduire une classifieur issu de la résolution de (P1).

(5) Les conditions KKT

Si (P1) est convexe et les conditions de qualification de Slater sont vérifiées.

Alors (h^*, ξ^*) est solution de (P1) et (α^*, β^*) est solution de (P2), ssi

① $\forall i \in [\![1,n]\!]$, $g_i(h,\xi) \leqslant 0$, $m_i(\xi) \leq 0$

② $\alpha^* \geqslant 0$

③ $\forall i \in [\![1,n]\!]$, $\alpha_i = 0$ ou $g_i(h,\xi) = 0$

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

Donc on a $h^* = \frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i^* \cdot Y_i \cdot k(\cdot, X_i)$

alors la classifieur est

$$\hat{g}^*(x) = \text{sign}(h^*(x))$$

6. (1½ points) Justifier que, connaissant $h^\star$, on peut choisir $\xi_i^\star = \max(0, 1 - Y_i h^\star(x_i))$ pour tout $i \in [\![1, n]\!]$. En déduire que pour tout $i \in [\![1, n]\!]$, si $Y_i h^\star(x_i) > 1$, alors $\alpha_i^\star = 0$.

(6) Si on connait h^*, d'après (P_1)

on a $\forall i \in [\![1,n]\!]$ $\begin{cases} \xi_i \geq 1 - Y_i h(X_i) \\ \xi_i \geq 0 \end{cases}$

donc on peut choisir $\xi_i^* = \max(0, 1 - Y_i h^*(x_i))$, $\forall i \in [\![1,n]\!]$

$\forall i \in [\![1,n]\!]$, si $Y_i h^*(x_i) > 1$, alors $\xi_i^* = 0$

et comme $\xi_i^* = \alpha_i^* + \beta_i^*$ et $\alpha_i^* \geq 0$, $\beta_i^* = 0$

alors $\alpha_i^* = 0$

7. (1 point (bonus)) Proposer un critère d'arrêt pour l'algorithme itératif de la question 4. Le justifier rapidement et expliciter le calcul.

(7) le critère d'arrêt est que tous les α_i sont vérifiées les conditions de KKT.

日期:　　　　/