Examen : Introduction à l'apprentissage automatique

18 décembre 2019

Tous les documents et les ordinateurs connectés sont autorisés.

Les questions peuvent être traitées de manière indépendante en admettant le résultat des questions précédentes.

Le barème (sur 19 points, auxquels s'ajoutent 4 points bonus) n'est donné qu'à titre indicatif.

Exercice 1 (Une variante des SVM, 5½ points)

Soient $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un échantillon de n couples aléatoires et $C \geq 0$. Pour tout $i \in [n]$, on appelle

$$\ell_i : x \in \mathbb{R} \mapsto \frac{Y_i + 3}{2} \max(0, 1 - Y_i x)$$

et on considère le problème d'optimisation :

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \ \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{1}{2} b^2 + C \sum_{i=1}^{n} \ell_i(w^\top X_i + b). \tag{P1}$$

1. a) (1 point) Tracer les graphes de ℓ_i pour $i \in [n]$ tel que $Y_i = 1$ et $Y_i = -1$. La résolution de (P1) permet-il de construire un régresseur ou un classifieur ? Expliquer votre réponse.

   Quelle est la différence entre le problème d'intérêt et une machine à vecteurs supports telle que vue en cours ?

   b) (1 point) Donner la forme de la fonction de *prédiction* en fonction d'un couple $(\hat{w}, \hat{b})$ solution de (P1). Est-ce un prédicteur linéaire ou non-linéaire ? Réaliser une figure illustrant le problème en jeu et la prédiction.

2. (1½ points) En notant $p = (\mathbf{1}_{Y_1=1}, \dots, \mathbf{1}_{Y_n=1}) \in \mathbb{R}^n$, $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, $A = \begin{pmatrix} Y_1 X_1^\top \\ \vdots \\ Y_n X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ et $\mathbb{1}$ le vecteur rempli de 1 (de taille adéquate), montrer que (P1) peut se réécrire

$$\begin{aligned} &\underset{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{1}{2} b^2 + C \mathbb{1}^\top \xi + C p^\top \xi \\ &\text{s. t.} \quad \begin{cases} Aw + by \succcurlyeq \mathbb{1} - \xi \\ \xi \succcurlyeq 0. \end{cases} \end{aligned} \tag{P2}$$

3. (2 points) Établir le problème d'optimisation dual à (P2).

Exercice 2 (Régression à noyau, 3 points)

Soient $\mathcal{H}$ un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ un échantillon de n couples aléatoires. On appelle $K = (k(X_i, X_j))_{1 \leq i,j \leq n}$ la matrice noyau, supposée inversible.

Pour $\alpha > 0$, on considère le régresseur $\hat{f}$ défini par :

$$\{\hat{f}\} = \arg\min_{f \in \mathcal{H}} \left\{ \alpha \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^n (Y_i - f(X_i))^2 \right\}.$$

1. (2 points) En faisant appel aux résultats du cours, justifier que $\hat{f} = \sum_{i=1}^n \hat{\beta}_i k(\cdot, X_i)$, où $\hat{\beta} \in \mathbb{R}^n$ est solution de
$$\arg\min_{\beta \in \mathbb{R}^n} \left\{ \alpha \beta^\top K \beta + \|y - K\beta\|_{\ell_2}^2 \right\},$$
où $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$.
2. (1 point) Montrer que $\hat{\beta} = (K + \alpha I_n)^{-1} y$, où $I_n \in \mathbb{R}^{n \times n}$ est la matrice identité.

Exercice 3 (Approximation de la matrice noyau, 10½ points)

Soient $\mathcal{H}$ un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ un ensemble de n points fixés et $m \leq n$ un entier non-nul.

On souhaite analyser quelques propriétés d'une méthode d'approximation de la matrice noyau $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$ par les m premiers points x_i. Pour ce faire, notons $K^m = (k(x_i, x_j))_{1 \leq i,j \leq m}$ la matrice noyau de $\{x_1, \dots, x_m\}$ (supposée inversible) et $P_m : \mathbb{R}^d \to \mathbb{R}^d$ le projecteur orthogonal sur $\mathcal{V} = \operatorname{span}\{k(\cdot, x_1), \dots, k(\cdot, x_m)\}$.

Par ailleurs, pour tout $f \in \mathcal{H}$, on appelle *coordonnées de $P_m f$ dans $\mathcal{V}$* un vecteur $\alpha \in \mathbb{R}^m$ vérifiant $P_m f = \sum_{\ell=1}^m \alpha_\ell k(\cdot, x_\ell)$.

1. a) (1 point) Soit $f \in \mathcal{H}$. Montrer que les coordonnées de $P_m f$ dans $\mathcal{V}$ s'expriment par $\alpha = (K^m)^{-1} F$, où $F \in \mathbb{R}^m$ est un vecteur à déterminer.

   b) (1 point) On appelle, pour tout $i, j \in [n]$, $f_i = k(\cdot, x_i)$ et $\tilde{K}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$. Expliquer en quoi $\tilde{K}_{i,j}$ peut être vu comme une approximation de $K_{i,j}$ et justifier que la matrice $\tilde{K} \in \mathbb{R}^{n \times n}$ est symétrique semi-définie positive.

   c) (1 point) Montrer que $\sum_{i=1}^n \|f_i - P_m f_i\|_{\mathcal{H}}^2 = \operatorname{tr}(K) - \operatorname{tr}(\tilde{K})$.

   d) (1 point) Montrer que pour tout $i, j \in [n]$, $K_{i,j} - \tilde{K}_{i,j} = \langle f_i - P_m f_i, f_j - P_m f_j \rangle_{\mathcal{H}}$. En déduire que $K - \tilde{K}$ est symétrique semi-définie positive.

2. a) (1 point) Pour tout $i \in [n]$, on appelle $\alpha_i \in \mathbb{R}^m$ le vecteur des coordonnées de $P_m f_i$ dans $\mathcal{V}$. Exprimer de manière matricielle $\tilde{K}_{i,j}$ ($i, j \in [n]$) en fonction de α_i et α_j.

b) (1 point) Soit $\Lambda = (\alpha_1|\dots|\alpha_n) \in \mathbb{R}^{m\times n}$. En identifiant la matrice K par blocs : $K = \begin{pmatrix} K^m & Q \\ Q^\top & K' \end{pmatrix}$, exprimer Λ et en déduire que $\tilde{K} = (K^m|Q)^\top (K^m)^{-1}(K^m|Q)$.

3. a) (½ point) Comparer l'espace mémoire nécessaire pour stocker K et $\tilde{K}$ (tel qu'exprimé à la question 2. b).

b) (1 point) Montrer que $\left\|K - \tilde{K}\right\|_F = \left\|K' - Q^\top (K^m)^{-1} Q\right\|_F$, où $\|\cdot\|_F$ est la norme de Frobenius.

c) (1 point) Soit $X \in \mathbb{R}^{n\times d}$ la matrice des données (rangées en ligne). En identifiant X par blocs : $X = \begin{pmatrix} X^m \\ X' \end{pmatrix}$, avec $X^m \in \mathbb{R}^{m\times d}$ la matrice des m premiers points, supposée de rang plein, on appelle $P_\top : \mathbb{R}^d \to \mathbb{R}^d$ le projecteur orthogonal sur l'espace engendré par les colonnes de $(X^m)^\top$ (c'est-à-dire les lignes de X^m). Montrer que $P_\top = (X^m)^\top (X^m (X^m)^\top)^{-1} X^m$.

d) (2 points) On considère (dans cette question uniquement) que k est le noyau linéaire, autrement dit $K = XX^\top$. Déduire de la question précédente que $\tilde{K} = XP_\top X^\top$.

4. a) (1 point (bonus)) En nommant $K = RDR^\top$ la décomposition en éléments propres de K ($D \in \mathbb{R}^{n\times n}$ est une matrice diagonale, $R \in \mathbb{R}^{n\times n}$ une matrice orthogonale) et $R = \begin{pmatrix} R^m \\ R' \end{pmatrix}$ la décomposition par blocs de R avec $R^m \in \mathbb{R}^{m\times n}$, exprimer K en fonction de D, R^m et R'.

b) (1 point (bonus)) On suppose que $\mathrm{rank}(K) = \mathrm{rank}(K^m) = m$. Que peut-on en déduire sur D ? Exprimer $(K^m)^{-1}$ en fonction de blocs de D et de R^m.

c) (2 points (bonus)) En déduire que lorsque $\mathrm{rank}(K) = \mathrm{rank}(K^m) = m$, $\tilde{K} = K$.

Exercice 1 (Une variante des SVM, 5½ points)

Soient $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un échantillon de n couples aléatoires et $C \geq 0$. Pour tout $i \in [n]$, on appelle

$$\ell_i : x \in \mathbb{R} \mapsto \frac{Y_i + 3}{2} \max(0, 1 - Y_i x)$$

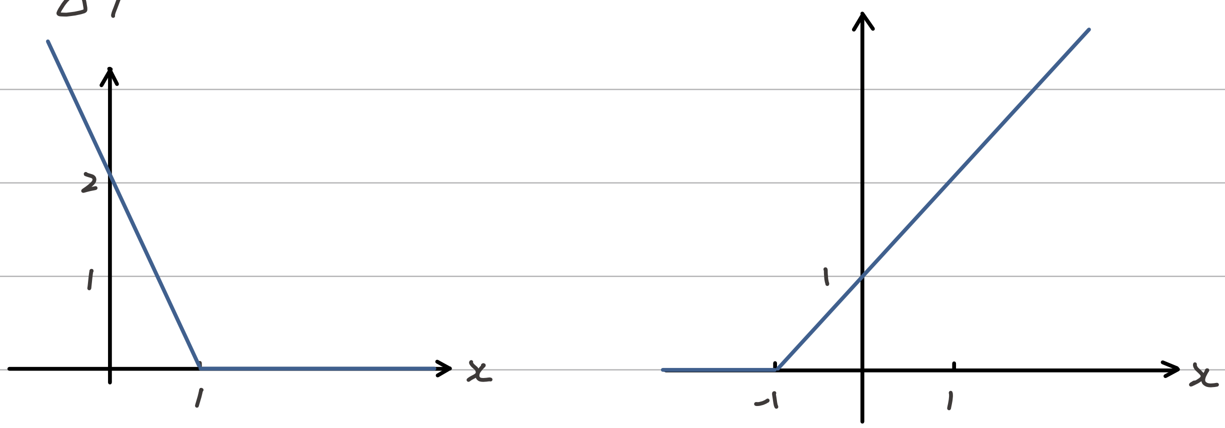
et on considère le problème d'optimisation :

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \ \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{1}{2} b^2 + C \sum_{i=1}^{n} \ell_i(w^\top X_i + b). \tag{P1}$$

1. a) (1 point) Tracer les graphes de ℓ_i pour $i \in [n]$ tel que $Y_i = 1$ et $Y_i = -1$. La résolution de (P1) permet-il de construire un régresseur ou un classifieur ? Expliquer votre réponse.
Quelle est la différence entre le problème d'intérêt et une machine à vecteurs supports telle que vue en cours ?

1) (a)

Les graphes de l_i



$Y_i = 1 \quad l_i(x) = 2 \cdot \max(0, 1-x)$ $\qquad$ $Y_i = -1 \quad l_i(x) = \max(0, 1+x)$

La solution de (P_1) permet de construire un classifieur.
La solution de (P_1) est (w^*, b^*), on peut construire un classifieur $g^*(x) = \text{sign}(w^* x + b^*)$

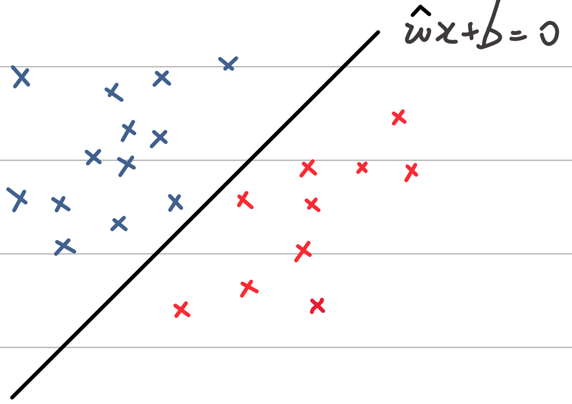
Pour le SVM en cours, $l_i^{\text{cours}}(x) = \max(0, 1 - Y_i x)$, la coefficient de loss dans le cas $Y_i = 1$ et $Y_i = -1$ sont le même.

b) (1 point) Donner la forme de la fonction de *prédiction* en fonction d'un couple $(\hat{w}, \hat{b})$ solution de (P1). Est-ce un prédicteur linéaire ou non-linéaire ? Réaliser une figure illustrant le problème en jeu et la prédiction.

(b) La fonction de prédiction est

$$\hat{g}(x) = \text{sign}(\hat{w}^\top x + b)$$

C'est un prédicteur linéaire.



2. (1½ points) En notant $p = (\mathbf{1}_{Y_1=1}, \ldots, \mathbf{1}_{Y_n=1}) \in \mathbb{R}^n$, $y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, $A = \begin{pmatrix} Y_1 X_1^\top \\ \vdots \\ Y_n X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ et $\mathbb{1}$ le vecteur rempli de 1 (de taille adéquate), montrer que (P1) peut se réécrire

$$\begin{array}{ll} \underset{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + C\mathbb{1}^\top\xi + Cp^\top\xi \\ \text{s. t.} & \begin{cases} Aw + by \succcurlyeq \mathbb{1} - \xi \\ \xi \succcurlyeq 0. \end{cases} \end{array} \tag{P2}$$

(2) $\forall i \in [\![1,n]\!]$, $l_i(w^\top X_i + b) = \frac{Y_i+3}{2} \cdot \max(0, 1 - Y_i(w^\top X_i + b))$

$$= \frac{Y_i+3}{2} \cdot \inf_{\substack{\xi_i \geq 0 \\ \xi_i \geq 1 - Y_i(w^\top X_i + b)}} \xi_i$$

Donc (P1) est équivalent à

$$\underset{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + C \cdot \sum_{i=1}^n \frac{Y_i+3}{2}\xi_i$$

$$\text{s.c. } \forall i \in [\![1,n]\!], \begin{cases} w^\top Y_i X_i + bY_i \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

et on a

$$\frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + C\cdot\sum_{i=1}^{n}\frac{Y_i+3}{2}\xi_i$$

$$= \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + 2C\cdot\sum_{i=1}^{n}\mathbb{1}_{\{Y_i=1\}}\cdot\xi_i + C\sum_{i=1}^{n}\mathbb{1}_{\{Y_i=-1\}}\cdot\xi_i$$

$$= \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + C\cdot\sum_{i=1}^{n}\mathbb{1}_{\{Y_i=1\}}\cdot\xi_i + C\cdot\sum_{i=1}^{n}\xi_i$$

$$= \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + Cp^T\xi + C\mathbb{1}^T\xi$$

Donc (P_1) est équivalent à

$$\underset{w\in\mathbb{R}^n,\, b\in\mathbb{R},\, \xi\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + Cp^T\xi + C\mathbb{1}^T\xi$$

$$\text{s.c.} \begin{cases} Aw + by \geqslant 1 - \xi \\ \xi \geqslant 0 \end{cases}$$

3. (2 points) Établir le problème d'optimisation dual à (P2).

(3) D'après (2), pour $\alpha\in\mathbb{R}_+^n$, $\beta\in\mathbb{R}_+^n$, la Lagrancien de (P_2) est

$$\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|_{\ell_2}^2 + \frac{1}{2}b^2 + Cp^T\xi + C\mathbb{1}^T\xi + \sum_{i=1}^{n}\alpha_i\cdot\left(1-\xi_i - Y_i(w^TX_i+b)\right) - \sum_{i=1}^{n}\beta_i\cdot\xi_i$$

$$\nabla_w\mathcal{L} = 0 \implies \hat{w} = \sum_{i=1}^{n}\alpha_i\cdot Y_i\cdot X_i$$

$$\nabla_b\mathcal{L} = 0 \implies \hat{b} = \sum_{i=1}^{n}\alpha_i Y_i$$

$$\nabla_\xi\mathcal{L} = 0 \implies \beta = C\cdot(p+\mathbb{1}) - \alpha$$

Donc

$$\begin{aligned}
\mathcal{L}(\hat{w}, \hat{b}, \hat{\xi}, \alpha, \beta) &= \frac{1}{2} \Big\| \sum_{i=1}^{n} \alpha_i Y_i X_i \Big\|_{\ell_2}^2 + \frac{1}{2} \Big(\sum_{i=1}^{n} \alpha_i Y_i \Big)^2 \\
&\quad + \sum_{i=1}^{n} \alpha_i \cdot \Big(1 - Y_i \cdot \Big(\sum_{j=1}^{n} \alpha_j Y_j X_j^T X_i + \sum_{i=1}^{n} \alpha_i Y_i \Big) \Big) \\
&= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j X_i X_j + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j \\
&\quad + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j X_i X_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j \\
&= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j (X_i X_j + 1) + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2} \alpha^T Q \alpha + \alpha^T \mathbb{1}
\end{aligned}$$

alors

$$D(\alpha) = \begin{cases} -\frac{1}{2} \alpha^T Q \alpha + \alpha^T \mathbb{1} & \text{si } 0 \leqslant \alpha \leqslant C \cdot (p + \mathbb{1}) \\ -\infty & \text{sinon} \end{cases}$$

donc le problème d'optimisation dual à (P_2) est

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

$$\text{s.c.} \left\{ 0 \leqslant \alpha \leqslant C \cdot (p + \mathbb{1}) \right\}$$

Exercice 2 (Régression à noyau, 3 points)

Soient $\mathcal{H}$ un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ un échantillon de n couples aléatoires. On appelle $K = (k(X_i, X_j))_{1\le i,j\le n}$ la matrice noyau, supposée inversible.

Pour $\alpha > 0$, on considère le régresseur $\hat{f}$ défini par :

$$\{\hat{f}\} = \arg\min_{f\in\mathcal{H}} \left\{ \alpha \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}.$$

1. (2 points) En faisant appel aux résultats du cours, justifier que $\hat{f} = \sum_{i=1}^n \hat{\beta}_i k(\cdot, X_i)$, où $\hat{\beta} \in \mathbb{R}^n$ est solution de

$$\arg\min_{\beta\in\mathbb{R}^n} \left\{ \alpha\beta^\top K\beta + \|y - K\beta\|_{\ell_2}^2 \right\},$$

où $y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$.

(1) pour $\psi : x \in \mathbb{R}_+ \longrightarrow \alpha x^2$ est un fonction croissante.

et $l : (x, y) \in \mathbb{R}^{2n} \longrightarrow \sum_{i=1}^{n} (x_i - y_i)^2$ un fonction de perte

D'après la théorème de représenter, la probléme d'optimisation

$$\underset{f\in\mathcal{H}}{\text{minimize}}\ \psi(\|f\|_{\mathcal{H}}) + l(Y_1, \ldots, Y_n, f(X_1), \ldots, f(X_n))$$

a une solution $\hat{f} = \sum_{i=1}^{n} \beta_i \cdot k(\cdot, X_i)$, $\beta \in \mathbb{R}^n$

Alors la probléme est équivalent à

$$\underset{\beta\in\mathbb{R}^n}{\text{minimize}}\ \alpha \Big\| \sum_{i=1}^{n} \beta_i k(\cdot, X_i) \Big\|_{\mathcal{H}}^2 + \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{n} \beta_j \cdot k(X_i, X_j) \Big)^2$$

$$\Longleftrightarrow$$

$$\underset{\beta\in\mathbb{R}^n}{\text{minimize}}\ \alpha \beta^T K \beta + \|y - K\beta\|_{\ell_2}^2$$

Donc $\hat{\beta} \in \mathbb{R}^n$ est la solution de

$$\underset{\beta\in\mathbb{R}^n}{\text{argmin}} \left\{ \alpha \beta^T K \beta + \|y - K\beta\|_{\ell_2}^2 \right\}$$

2. (1 point) Montrer que $\hat{\beta} = (K + \alpha I_n)^{-1} y$, où $I_n \in \mathbb{R}^{n \times n}$ est la matrice identité.

(2) Pour le probléme

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \ \alpha \beta^T K \beta + \| y - K\beta \|_{\ell_2}^2$$

soit $\varphi(\beta) = \alpha \beta^T K \beta + y^T y + \beta^T K^T K \beta - 2 \beta^T K^T y$

$$\nabla_\beta \varphi = 2\alpha K \beta + 2 K^T K \beta - 2 K^T y$$

on prend $\nabla_\beta \varphi = 0$

$$\alpha K \beta + K^T K \beta = K^T y$$

$$\hat{\beta} = (K + \alpha I_n)^{-1} y$$

Exercice 3 (Approximation de la matrice noyau, 10½ points)

Soient $\mathcal{H}$ un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ un ensemble de n points fixés et $m \leq n$ un entier non-nul.

On souhaite analyser quelques propriétés d'une méthode d'approximation de la matrice noyau $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$ par les m premiers points x_i. Pour ce faire, notons $K^m = (k(x_i, x_j))_{1 \leq i,j \leq m}$ la matrice noyau de $\{x_1, \ldots, x_m\}$ (supposée inversible) et $P_m : \mathbb{R}^d \to \mathbb{R}^d$ le projecteur orthogonal sur $\mathcal{V} = \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_m)\}$.

Par ailleurs, pour tout $f \in \mathcal{H}$, on appelle *coordonnées de $P_m f$ dans $\mathcal{V}$* un vecteur $\alpha \in \mathbb{R}^m$ vérifiant $P_m f = \sum_{\ell=1}^m \alpha_\ell k(\cdot, x_\ell)$.

1. a) (1 point) Soit $f \in \mathcal{H}$. Montrer que les coordonnées de $P_m f$ dans $\mathcal{V}$ s'expriment par $\alpha = (K^m)^{-1} F$, où $F \in \mathbb{R}^m$ est un vecteur à déterminer.

1) (a) Pour les m points non-null $(x_1, \ldots, x_m)$

$\forall i \in [\![1, m]\!]$, $\quad P_m f(x_i) = \sum_{\ell=1}^m \alpha_\ell \cdot k(x_i, x_\ell)$

alors

$$\begin{bmatrix} P_m f(x_1) \\ \vdots \\ P_m f(x_m) \end{bmatrix} = \begin{bmatrix} \sum_{\ell=1}^m \alpha_\ell \cdot k(x_1, x_\ell) \\ \vdots \\ \sum_{\ell=1}^m \alpha_\ell \cdot k(x_m, x_\ell) \end{bmatrix} = K^m \alpha$$

donc

$$\alpha = (K^m)^{-1} F \qquad \text{avec } F = \begin{bmatrix} P_m f(x_1) \\ \vdots \\ P_m f(x_m) \end{bmatrix}$$

b) (1 point) On appelle, pour tout $i, j \in [n]$, $f_i = k(\cdot, x_i)$ et $\tilde{K}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$. Expliquer en quoi $\tilde{K}_{i,j}$ peut être vu comme une approximation de $K_{i,j}$ et justifier que la matrice $\tilde{K} \in \mathbb{R}^{n \times n}$ est symétrique semi-définie positive.

(b) $\forall i, j \in [\![1, m]\!]$, $f_i = k(\cdot, x_i) \in V$, $P_m f_i = f_i = k(\cdot, x_i)$

alors $\tilde{k}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} = \langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}}$

$= k(x_i, x_j) = K_{ij}$

pour $i \in [\![1, m]\!]$, $\forall j \in [\![m+1, n]\!]$

on $P_m f_i = k(\cdot, x_i)$, $\quad P_m f_j = \sum_{\ell=1}^m \alpha_\ell^{(j)} \cdot k(\cdot, x_\ell)$

alors $\tilde{k}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} = \langle k(\cdot, x_i), \sum_{\ell=1}^m \alpha_\ell^{(j)} \cdot k(\cdot, x_\ell) \rangle_{\mathcal{H}}$

$= \sum_{\ell=1}^m \alpha_\ell^{(j)} \cdot k(x_i, x_\ell) = P_m f_j(x_i)$

日期:　　/

donc $\begin{bmatrix} \tilde{k}_{1,j} \\ \vdots \\ \tilde{k}_{m,j} \end{bmatrix} = \begin{bmatrix} \sum_{l=1}^{m} \alpha_l^{(j)} \cdot k(x_1, x_l) \\ \vdots \\ \sum_{l=1}^{m} \alpha_l^{(j)} \cdot k(x_m, x_l) \end{bmatrix} = K^m \cdot (K^m)^{-1} \cdot F^{(j)} = \begin{bmatrix} P_m f_j(x_1) \\ \vdots \\ P_m f_j(x_m) \end{bmatrix}$

pour $i \in [\![m+1, n]\!]$, $j \in [\![m+1, n]\!]$,

$$P_m f_i = \sum_{l=1}^{m} \alpha_l^{(i)} \cdot k(\cdot, x_l) \qquad P_m f_j = \sum_{l=1}^{m} \alpha_l^{(j)} \cdot k(\cdot, x_l)$$

alors
$$\begin{aligned} \tilde{k}_{i,j} &= \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} \\ &= \langle \sum_{l=1}^{m} \alpha_l^{(i)} \cdot k(\cdot, x_l), \sum_{l=1}^{m} \alpha_l^{(j)} \cdot k(\cdot, x_l) \rangle_{\mathcal{H}} \\ &= \sum_{l=1}^{m} \sum_{t=1}^{m} \alpha_l^{(i)} \cdot \alpha_t^{(j)} \cdot k(x_t, x_l) \\ &= \alpha^{(i)T} K^m \alpha^{(j)} \\ &= ((K^m)^{-1} F^{(i)})^T K^m (K^m)^{-1} F^{(j)} \\ &= F^{(i)T} (K^m)^{-1^T} F^{(j)} \end{aligned}$$

———— (另一种思路)

① $\forall i \in [\![1, m]\!]$, $\forall j \in [\![1, m]\!]$

$$\tilde{k}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} = \langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}} = k(x_i, x_j) = k_{i,j}$$

② $\forall i \in [\![1, m]\!]$, $\forall j \in [\![m+1, n]\!]$

$$\tilde{k}_{ij} = \langle k(\cdot, x_i), \sum_{l=1}^{m} \alpha_{jl} \cdot k(\cdot, x_l) \rangle_{\mathcal{H}} = \sum_{l=1}^{m} \alpha_{jl} k(x_i, x_l) = P_m f_j(x_i)$$

comme $f_j(x_i) = k(x_i, x_j) = k(x_j, x_i) = f_i(x_j)$ et $f_i \in V$

alors $P_m f_j(x_i) = P_m f_i(x_j) = f_i(x_j) = k(x_j, x_i) = k(x_i, x_j)$

Donc $\tilde{k}_{ij} = k_{ij}$

③ $\forall i \in [\![m+1, n]\!]$, $\forall j \in [\![1, m]\!]$

la même $\tilde{k}_{ij} = k_{ij}$

日期: /

④ $\forall i \in [\![m+1, n]\!]$, $\forall j \in [\![m+1, n]\!]$

$$\tilde{K}_{ij} = \langle P_m f_i, P_m f_j \rangle_H = \langle \sum_{l=1}^{m} \alpha_{il} \cdot k(\cdot, x_l), \sum_{t=1}^{m} \alpha_{jt} \cdot k(\cdot, x_t) \rangle_H$$

$$= \sum_{l=1}^{m} \sum_{t=1}^{m} \alpha_{il} \cdot \alpha_{jt} \cdot k(x_t, x_l)$$

$$= \alpha_i^T K^m \alpha_j$$

Donc $\tilde{K}_{ij}$ peut être vue comme une approximation de K_{ij}.

Et on a $\forall i,j \in [\![1, n]\!]$, $\tilde{K}_{ij} = \tilde{K}_{ji}$

et comme K est semi-définie positive,
alors $\tilde{K}$ est semi-définie positive

c) (1 point) Montrer que $\sum_{i=1}^{n} \|f_i - P_m f_i\|_{\mathcal{H}}^2 = \mathrm{tr}(K) - \mathrm{tr}(\tilde{K})$.

(c) Pour $i \in [\![1, m]\!]$,

$$\sum_{i=1}^{m} \|f_i - P_m f_i\|_{\mathcal{H}}^2 = \sum_{i=1}^{m} \|f_i - f_i\|_{\mathcal{H}}^2 = 0$$

donc il faut montrer

$$\sum_{i=m+1}^{n} \|f_i - P_m f_i\|_{\mathcal{H}}^2 = tr(K) - tr(\tilde{K})$$

alors

$$\begin{aligned} tr(K) - tr(\tilde{K}) &= tr(K - \tilde{K}) \\ &= 0 + \sum_{i=m+1}^{n} \left(k(x_i, x_i) - \langle P_m f_i, P_m f_i \rangle_{\mathcal{H}} \right) \\ &= \sum_{i=m+1}^{n} \left(k(x_i, x_i) - \alpha^{(i)T} K^m \alpha^{(i)} \right) \end{aligned}$$

$\forall i \in [\![m+1, n]\!]$

$$\begin{aligned} \|f_i - P_m f_i\|_{\mathcal{H}}^2 &= \langle f_i - P_m f_i, f_i - P_m f_i \rangle_{\mathcal{H}} \\ &= \langle f_i, f_i \rangle_{\mathcal{H}} + \langle P_m f_i, P_m f_i \rangle_{\mathcal{H}} - 2 \langle f_i, P_m f_i \rangle_{\mathcal{H}} \\ &= k(x_i, x_i) + \langle P_m f_i, P_m f_i \rangle_{\mathcal{H}} \\ &\quad - 2 \langle P_m f_i + P_m^{\perp} f_i, P_m f_i \rangle_{\mathcal{H}} \\ &= k(x_i, x_i) - \langle P_m f_i, P_m f_i \rangle_{\mathcal{H}} \end{aligned}$$

Alors

$$\sum_{i=1}^{n} \|f_i - P_m f_i\|_{\mathcal{H}}^2 = tr(K) - tr(\tilde{K})$$

d) (1 point) Montrer que pour tout $i, j \in [n]$, $K_{i,j} - \tilde{K}_{i,j} = \langle f_i - P_m f_i, f_j - P_m f_j \rangle_{\mathcal{H}}$. En déduire que $K - \tilde{K}$ est symétrique semi-définie positive.

(d)

$$K_{i,j} - \tilde{K}_{i,j} = k(x_i, x_j) - \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$$
$$= \langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}} - \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$$
$$= \langle f_i, f_j \rangle_{\mathcal{H}} - \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$$

$$\langle f_i - P_m f_i, f_j - P_m f_j \rangle_{\mathcal{H}}$$
$$= \langle f_i, f_j \rangle_{\mathcal{H}} + \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} - \langle f_i, P_m f_j \rangle_{\mathcal{H}} - \langle P_m f_i, f_j \rangle_{\mathcal{H}}$$
$$= \langle f_i, f_j \rangle_{\mathcal{H}} + \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}} - \langle P_m f_i + P_m^{\perp} f_i, P_m f_j \rangle_{\mathcal{H}}$$
$$- \langle P_m f_i, P_m f_j + P_m^{\perp} f_j \rangle_{\mathcal{H}}$$
$$= \langle f_i, f_j \rangle_{\mathcal{H}} - \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$$

Donc $K_{i,j} - \tilde{K}_{i,j} = \langle f_i - P_m f_i, f_j - P_m f_j \rangle_{\mathcal{H}}$

Donc $(K - \tilde{K})_{i,j} = (K - \tilde{K})_{j,i}$ alors $K - \tilde{K}$ est symétrique

D'après (c) $tr(K - \tilde{K}) \geq 0$

et comme K et $\tilde{K}$ sont semi-définie positive

alors $K - \tilde{K}$ est semi-définie positive

2. a) (1 point) Pour tout $i \in [n]$, on appelle $\alpha_i \in \mathbb{R}^m$ le vecteur des coordonnées de $P_m f_i$ dans $\mathcal{V}$. Exprimer de manière matricielle $\tilde{K}_{i,j}$ ($i, j \in [n]$) en fonction de α_i et α_j.

(2) (a)

$$\begin{aligned}
\tilde{K}_{i,j} &= \langle P_m f_i , P_m f_j \rangle_{\mathcal{H}} \\
&= \langle \sum_{\ell=1}^{m} \alpha_{i\ell} \cdot k(\cdot, x_\ell) , \sum_{\ell=1}^{m} \alpha_{j\ell} \cdot k(\cdot, x_\ell) \rangle_{\mathcal{H}} \\
&= \sum_{\ell=1}^{m} \sum_{t=1}^{m} \alpha_{i\ell} \cdot \alpha_{jt} \cdot k(x_\ell, x_t) \\
&= \alpha_i^{\top} K^m \alpha_j
\end{aligned}$$

b) (1 point) Soit $\Lambda = (\alpha_1 | \ldots | \alpha_n) \in \mathbb{R}^{m \times n}$. En identifiant la matrice K par blocs : $K = \begin{pmatrix} K^m & Q \\ Q^\top & K' \end{pmatrix}$, exprimer Λ et en déduire que $\tilde{K} = (K^m|Q)^\top (K^m)^{-1} (K^m|Q)$.

(b) $\forall i \in [\![1, n]\!]$, $\alpha_i = (K^m)^{-1} F_i$

on a

$$\begin{aligned}
\Lambda &= (K^m)^{-1} \left[F_1 | \cdots | F_n \right] \\
&= (K^m)^{-1} \begin{bmatrix} P_m f_1(x_1) & P_m f_2(x_1) & \cdots & P_m f_m(x_1) & \cdots & P_m f_n(x_1) \\ P_m f_1(x_2) & P_m f_2(x_2) & \cdots & P_m f_m(x_2) & \cdots & P_m f_n(x_2) \\ \vdots & \vdots & & \vdots & & \vdots \\ P_m f_1(x_m) & P_m f_2(x_m) & \cdots & P_m f_m(x_m) & \cdots & P_m f_n(x_m) \end{bmatrix} \\
&= (K^m)^{-1} \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_m) & \cdots & P_m f_n(x_1) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_m) & \cdots & P_m f_n(x_2) \\ \vdots & \vdots & & \vdots & & \vdots \\ k(x_m, x_1) & k(x_m, x_2) & \cdots & k(x_m, x_m) & \cdots & P_m f_n(x_m) \end{bmatrix} \\
&= (K^m)^{-1} \left[\begin{array}{c|ccc} & P_m f_{m+1}(x_1) & \cdots & P_m f_n(x_1) \\ K^m & \vdots & & \vdots \\ & P_m f_{m+1}(x_m) & \cdots & P_m f_n(x_m) \end{array} \right] \\
&= (K^m)^{-1} \left[K^m \mid Q \right]
\end{aligned}$$

日期: /

D'après (a) $\tilde{K} = \Lambda^T K^m \Lambda$

$$= [K^m | Q]^T (K^m)^{-1} K^m (K^m)^{-1} [K^m | Q]$$

$$= [K^m | Q]^T (K^m)^{-1} [K^m | Q]$$

3. a) (½ point) Comparer l'espace mémoire nécessaire pour stocker K et $\tilde{K}$ (tel qu'exprimé à la question 2. b).

(3) (a) Pour K, l'espace est $\frac{n \times n}{2}$

Pour $\tilde{K}$, l'espace est $\frac{m \times m}{2} + m \times (n-m)$

b) (1 point) Montrer que $\left\| K - \tilde{K} \right\|_F = \| K' - Q^\top (K^m)^{-1} Q \|_F$, où $\|\cdot\|_F$ est la norme de Frobenius.

(b)

$$\| K - \tilde{K} \|_F = \sqrt{tr[(K-\tilde{K})^T (K-\tilde{K})]}$$

$$= \sqrt{tr(K-\tilde{K})^T} \cdot \sqrt{tr(K-\tilde{K})}$$

D'après (1) (c)

$$tr(K-\tilde{K}) = \sum_{i=m+1}^{n} \left(k(x_i, x_i) - \alpha_i^T K^m \alpha_i \right)$$

$$= tr(K' - Q^T (K^m)^{-1} Q)$$

donc

$$\| K - \tilde{K} \|_F = \sqrt{tr(K' - Q^T (K^m)^{-1} Q)^T} \sqrt{tr(K' - Q^T (K^m)^{-1} Q)}$$

$$= \| K' - Q^T (K^m)^{-1} Q \|_F$$

c) (1 point) Soit $X \in \mathbb{R}^{n\times d}$ la matrice des données (rangées en ligne). En identifiant X par blocs : $X = \begin{pmatrix} X^m \\ X' \end{pmatrix}$, avec $X^m \in \mathbb{R}^{m\times d}$ la matrice des m premiers points, supposée de rang plein, on appelle $P_\top : \mathbb{R}^d \to \mathbb{R}^d$ le projecteur orthogonal sur l'espace engendré par les colonnes de $(X^m)^\top$ (c'est-à-dire les lignes de X^m). Montrer que $P_\top = (X^m)^\top (X^m (X^m)^\top)^{-1} X^m$.

(c) $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ alors $X^m = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}$ $(X^m)^T = (x_1, \dots, x_m)$

alors l'espace engendré par les colonnes de $(X^m)^T$ est $V_X = \text{span}\{x_1, \dots, x_m\} \subset \mathbb{R}^d$

$\forall\, Y \in \mathbb{R}^d$, soit $\hat{Y} = P_\perp Y \in V_X$,
alors il existe $\hat{\beta} \in \mathbb{R}^d$, $\hat{Y} = (X^m)^T \hat{\beta}$, et on a

$$\begin{cases} \langle x_1, Y - (X^m)^T \hat{\beta} \rangle_{\ell_2} = 0 \\ \vdots \\ \langle x_m, Y - (X^m)^T \hat{\beta} \rangle_{\ell_2} = 0 \end{cases}$$

alors on a

$$X^m (Y - (X^m)^T \hat{\beta}) = 0$$
$$X^m (X^m)^T \hat{\beta} = X^m Y$$
$$\hat{\beta} = (X^m (X^m)^T)^{-1} X^m Y$$

donc $\hat{Y} = (X^m)^T \hat{\beta} = (X^m)^T (X^m (X^m)^T)^{-1} X^m Y = P_T Y$

et on a

$$P_T = (X^m)^T (X^m (X^m)^T)^{-1} X^m$$

d) (2 points) On considère (dans cette question uniquement) que k est le noyau linéaire, autrement dit $K = XX^\top$. Déduire de la question précédente que $\tilde{K} = XP_\top X^\top$.

(d) D'après (2) (b)

$$K = \begin{bmatrix} K^m & Q \\ Q^\top & K' \end{bmatrix} \qquad \tilde{K} = \begin{bmatrix} K^m & Q \\ Q^\top & Q^\top (K^m)^{-1} Q \end{bmatrix}$$

et

$$\tilde{K} = (K^m | Q)^\top (K^m)^{-1} (K^m | Q)$$

$$P_\perp = (X^m)^\top (K^m)^{-1} X^m$$

alors

$$X P_\perp X^\perp = X (X^m)^\perp (K^m)^{-1} X^m X^\perp$$

$$= \begin{pmatrix} X^m (X^m)^\perp \\ X' (X^m)^\perp \end{pmatrix} (K^m)^{-1} \left(X^m (X^m)^\perp \,\middle|\, X' (X^m)^\perp \right)$$

$$= (K^m | Q)^\perp (K^m)^{-1} (K^m | Q)$$

$$= \tilde{K}$$

4. a) (1 point (bonus)) En nommant $K = RDR^\top$ la décomposition en éléments propres de K ($D \in \mathbb{R}^{n\times n}$ est une matrice diagonale, $R \in \mathbb{R}^{n\times n}$ une matrice orthogonale) et $R = \begin{pmatrix} R^m \\ R' \end{pmatrix}$ la décomposition par blocs de R avec $R^m \in \mathbb{R}^{m\times n}$, exprimer K en fonction de D, R^m et R'.

(4) (a)

$$K = RDR^T$$

$$= \begin{bmatrix} R^m \\ R' \end{bmatrix} D \left[(R^m)^T, (R')^T \right]$$

$$= \begin{bmatrix} R^m D (R^m)^T & R^m D (R')^T \\ (R^m D (R')^T)^T & R' D (R')^T \end{bmatrix}$$

b) (1 point (bonus)) On suppose que $\text{rank}(K) = \text{rank}(K^m) = m$. Que peut-on en déduire sur D? Exprimer $(K^m)^{-1}$ en fonction de blocs de D et de R^m.

(b) Si $\text{rank}(K) = \text{rank}(K^m) = m$, alors

$$D = \begin{bmatrix} d_{11} & & & & \\ & \ddots & & & \\ & & d_{mm} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix}_{n\times n}$$

donc

$$K^m = R^m D (R^m)^T$$

$$(K^m)^{-1} = (R^m D (R^m)^T)^{-1}$$

c) (2 points (bonus)) En déduire que lorsque $\text{rank}(K) = \text{rank}(K^m) = m$, $\tilde{K} = K$.

1c) Lorsque $\text{rank}(K) = \text{rank}(K^m) = m$

on a $(K^m)^{-1} = (R^m D (R^m)^T)^{-1}$

alors

$$\tilde{K}' = Q^T (K^m)^{-1} Q$$
$$= (R^m D (R')^T)^T (R^m D (R^m)^T)^{-1} (R^m D (R')^T)$$
$$= R' D (R^m)^T ((R^m)^T)^{-1} D^{-1} (R^m)^{-1} R^m D (R')^T$$
$$= R' D (R')^T$$
$$= K'$$

donc

$$K = \tilde{K}$$