Examen : Introduction à l'apprentissage automatique

12 novembre 2021

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 2 points bonus) n'est donné qu'à titre indicatif.

Exercice 1 (Questions de cours, 4 points)

1. (1 point) Soient $\{(x_i, y_i)\}_{1\le i\le n} \subset \mathbb{R}^d \times \{\pm 1\}$ et $C > 0$. Construire un classifieur SVM revient à déterminer

$$(\hat{w}_n, \hat{b}_n) \in \arg\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \ \frac{1}{2}\|w\|_{\ell_2}^2 + C\sum_{i=1}^n \max\left(0, 1 - y_i(w^\top x_i + b)\right).$$

   a) Expliquer le rôle de chacun des deux termes dans la fonction à minimiser.

   b) Quelle est la particularité de ce modèle par rapport à celui de régression logistique ?

2. (1 point) Soient $\{(x_i, y_i)\}_{1\le i\le n} \subset \mathbb{R}^d \times \mathbb{R}$, $\lambda > 0$ et $\mathcal{H}$ est RKHS de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. On définit alors

$$L : h \in \mathcal{H} \mapsto \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 + \sum_{i=1}^n \|h(x_i) - y_i\|_{\ell_2}^2.$$

   a) Pour tout $h \in \mathcal{H}$, on appelle $h_\parallel$ la projection de h sur $\mathrm{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\}$ et $h_\perp = h - h_\parallel$. Montrer que $\sum_{i=1}^n \|h(x_i) - y_i\|_{\ell_2}^2 = \sum_{i=1}^n \left\|h_\parallel(x_i) - y_i\right\|_{\ell_2}^2$.

   b) En déduire que $L(h) \ge L(h_\parallel)$.

3. (2 points) Soient $r > 0$ et $x \in \mathbb{R}^d$ tel que $\|x\|_{\ell_2} > r$. On souhaite retrouver la projection de x sur la boule de rayon r, $\mathcal{B}_r = \left\{y \in \mathbb{R}^d, \|y\|_{\ell_2} \le r\right\}$, par dualité lagrangienne. Pour ce faire, on résout

$$\begin{aligned} \underset{y\in\mathbb{R}^d}{\text{minimiser}} \quad & \|y - x\|_{\ell_2}^2 \\ \text{s. c.} \quad & \|y\|_{\ell_2}^2 \le r^2. \end{aligned}$$

   a) Vérifier que le problème et convexe et que les conditions de qualification de Slater s'appliquent.

   b) Définir un lagrangien pour ce problème.

   c) Énoncer les conditions KKT.

   d) En déduire une expression de la projection de x sur $\mathcal{B}_r$.

Exercice 2 (Algorithme EM, 7½ points)

Soient $\{(X_i, Y_i)\}_{1 \le i \le n}$ des paires indépendantes de variables aléatoires à valeurs dans $\mathbb{R} \times \{b_1, b_2\}$, avec $\{b_1, b_2\} \subset \mathbb{R}$, telles que pour tout $i \in [\![1, n]\!]$:

$$\mathbb{P}(Y_i = b_1) = \alpha_0 \qquad \text{et} \qquad X_i \,\middle|\, Y_i \sim \mathcal{N}\left(a_i^\top \beta_0 + Y_i, \sigma_0^2\right),$$

où $(\alpha_0, \beta_0, \sigma_0^2) \in]0,1[\times \mathbb{R}^d \times \mathbb{R}_+^*$ est un jeu de paramètres inconnus (les autres sont connus). On suppose ici que l'on n'observe que $\{X_1, \dots, X_n\}$ et l'on souhaite estimer $(\alpha_0, \beta_0, \sigma_0^2)$ par l'algorithme EM.

1. (1 point) Montrer que la formulation :
$$\begin{cases} X_i = a_i^\top \beta_0 + Y_i + \epsilon_i, \forall i \in [\![1, n]\!] \\ \epsilon \sim \mathcal{N}(0, \sigma_0^2 I_n) \\ \epsilon \perp\!\!\!\perp (Y_1, \dots, Y_n) \\ \{Y_1, \dots, Y_n\} \text{ i.i.d avec } \mathbb{P}(Y_1 = b_1) = \alpha_0 \end{cases}$$
est compatible avec le modèle posé (en particulier, on pourra utiliser que deux couples (X_i, Y_i) et (X_j, Y_j), pour $i \neq j$ dans $[\![1, n]\!]$, sont indépendants si pour toutes fonctions boréliennes bornées φ et ψ, $\mathbb{E}[\varphi(X_i, Y_i)\psi(X_j, Y_j)] = \mathbb{E}[\varphi(X_i, Y_i)]\,\mathbb{E}[\psi(X_j, Y_j)]$).

2. (1 point) En déduire une interprétation dudit modèle (on pourra se placer dans le cas $d = 1$ et proposer une représentation graphique).
Le modèle correspond-il à un problème de classification ou de régression ? À plan d'expérience fixé (*fixed design*) ou aléatoire ?

3. (1 point) Donner la loi jointe de (X_1, Y_1) (on précisera une mesure dominante). En déduire un modèle statistique pour la loi de (X_1, Y_1) puis l'expression de la log-vraisemblance $\ell_{(X,Y)_1^n}(\alpha, \beta, \sigma^2)$ d'un paramètre quelconque $(\alpha, \beta, \sigma^2) \in]0,1[\times \mathbb{R}^d \times \mathbb{R}_+^*$ (au regard de $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$).

4. (1 point) Expliciter, pour tout $x \in \mathbb{R}$, la loi de $Y_1 \,|\, X_1 = x$.
On suppose disposer d'un estimateur candidat $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ de $(\alpha_0, \beta_0, \sigma_0^2)$. Construire n variables aléatoires $Z_1, \dots, Z_n$ visant à « approcher » $Y_1, \dots, Y_n$, connaissant l'estimateur $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$.

5. (1 point) En déduire que pour tout $(\alpha, \beta, \sigma^2) \in]0,1[\times \mathbb{R}^d \times \mathbb{R}_+^*$,
$$\begin{aligned} &\mathbb{E}[\ell_{(X,Z)_1^n}(\alpha, \beta, \sigma^2) \,|\, X_1, \dots, X_n] \\ &= \log(\alpha) \sum_{i=1}^n p_i + \log(1-\alpha)\left(n - \sum_{i=1}^n p_i\right) - \frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) \\ &- \frac{1}{2\sigma^2} \sum_{i=1}^n \left[p_i \left(X_i - (a_i^\top \beta + b_1)\right)^2 + (1 - p_i)\left(X_i - (a_i^\top \beta + b_2)\right)^2\right], \end{aligned}$$
où $p_1, \dots, p_n$ sont à déterminer.

6. (2 points) En appelant $A = \begin{pmatrix} a_1^\top \\ \vdots \\ a_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ et en supposant que $\text{rang}(A) = d$, déterminer
$$\arg\max_{(\alpha, \beta, \sigma^2) \in]0,1[\times \mathbb{R}^d \times \mathbb{R}_+^*} \mathbb{E}[\ell_{(X,Z)_1^n}(\alpha, \beta, \sigma^2) \,|\, X_1, \dots, X_n].$$

7. (½ point) Décrire l'algorithme EM adapté au problème posé.

Exercice 3 (Clustering spectral, 8½ points)

Soient $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ et $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ une mesure de similarité symétrique. On appelle $W = (s(x_i, x_j))_{1 \le i,j \le n}$ la matrice d'adjacence des données, $D = \operatorname{diag}(d_1, \dots, d_n) \in \mathbb{R}^{n \times n}$ la matrice diagonale des degrés $d_i = \sum_{j=1}^n W_{ij}$ et $L = D - W$ le laplacien non-normalisé du graphe associé. Soient de plus $L_s = D^{-1/2} L D^{-1/2}$ et $L_w = D^{-1} L$ les laplaciens normalisés.

Préliminaires

1. (½ point) Donner une condition suffisante sur s pour que D soit non-singulière.
2. (½ point) On suppose D non-singulière. Montrer que $u \in \mathbb{R}^n$ est vecteur propre de L_w avec pour valeur propre $\lambda \in \mathbb{R}$ si est seulement si $D^{1/2} u$ est vecteur propre de L_s avec pour valeur propre λ.
3. (½ point) En déduire que $\mathbb{1}$ et $(\sqrt{d_1}, \dots, \sqrt{d_n})$ sont vecteurs propres de L_w et L_s respectivement et déterminer les valeurs propres associées.

Partie A

Pour une partie $I \subset [\![1, n]\!]$, on définit $\operatorname{vol}(I) = \sum_{i \in I} d_i$ et le vecteur

$$f_I = \left(\sqrt{\frac{\operatorname{vol}(I^c)}{\operatorname{vol}(I)}} \mathbf{1}_{i \in I} - \sqrt{\frac{\operatorname{vol}(I)}{\operatorname{vol}(I^c)}} \mathbf{1}_{i \in I^c} \right)_{1 \le i \le n},$$

où I^c est le complémentaire de I dans $[\![1, n]\!]$.

1. (1 point) Montrer que pour $i \in I$ et $j \in I^c$, en notant f_{Ii} la i^e composante de f_I,
$$(f_{Ii} - f_{Ij})^2 = \frac{\operatorname{tr}(D)}{\operatorname{vol}(I)} + \frac{\operatorname{tr}(D)}{\operatorname{vol}(I^c)}.$$
2. (1 point) Montrer que $\mathbb{1}^\top D f_I = 0$ et $f_I^\top D f_I = \operatorname{tr}(D)$.
3. (1 point) Sachant que pour tout $u \in \mathbb{R}^n$, $u^\top L u = \frac{1}{2} \sum_{1 \le i,j \le n} W_{ij} (u_i - u_j)^2$, montrer que
$$f_I^\top L f_I = \operatorname{tr}(D) \left(\frac{\sum_{\substack{i \in I \\ j \in I^c}} W_{ij}}{\operatorname{vol}(I)} + \frac{\sum_{\substack{i \in I^c \\ j \in I}} W_{ij}}{\operatorname{vol}(I^c)} \right).$$
4. (1 point) En déduire une réécriture du problème de *Normalized cut* dans le cas d'une seule coupure (*i.e.* d'un partitionnement en deux groupes) et un relâchement de celui-ci sous la forme d'un problème d'optimisation continue.
5. (1 point) Expliciter une solution du problème relâché.

Partie B

On s'intéresse à présent au problème de partitionnement en k (entier supérieur à 2) groupes par clustering spectral minimisant le coût *Normalized cut*, et on nomme donc $U \in \mathbb{R}^{n\times k}$ la matrice dont les colonnes sont les vecteurs propres de L_s associés aux k plus petites valeurs propres. On souhaite, à partir de cette matrice, remonter à une partition $\boldsymbol{I} = (I_1, \dots, I_k)$ de $[\![1, n]\!]$ telle que « le sous-espace vectoriel engendré par la partition $\boldsymbol{I}$ » soit aussi proche que possible de celui engendré par les colonnes de $H = D^{-1/2}U$, la matrice solution du problème relâché. Autrement dit, on souhaite avoir $\operatorname{range}(D^{1/2}Y_{\boldsymbol{I}}) \approx \operatorname{range}(U)$, où $Y_{\boldsymbol{I}} = (\mathbf{1}_{i\in I_j})_{\substack{1\le i\le n\\ 1\le j\le k}} \in \mathbb{R}^{n\times k}$ est la matrice *one-hot encoding* de la partition $\boldsymbol{I}$.

Pour ce faire, on utilise une distance entre projecteurs :

$$\mathcal{L}(\boldsymbol{I}) = \frac{1}{2}\|P_U - P_{\boldsymbol{I}}\|_F^2 = k - \sum_{j=1}^{k} \frac{1}{\operatorname{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell,$$

où P_U et $P_{\boldsymbol{I}}$ sont les projecteurs orthogonaux sur $\operatorname{range}(U)$ et $\operatorname{range}(D^{1/2}Y_{\boldsymbol{I}})$ respectivement, $u_i^\top$ est la i^e ligne de U et où l'on a remarqué que $\sum_{\substack{1\le i\le n\\ 1\le j\le k}} U_{ij}^2 = k$.

1. (2 points) Montrer que, pour tout $j \in [\![1, k]\!]$,

$$\min_{\mu\in\mathbb{R}^k} \sum_{i\in I_j} d_i \left\| \frac{u_i}{\sqrt{d_i}} - \mu \right\|_{\ell_2}^2 = \sum_{i\in I_j} \|u_i\|_{\ell_2}^2 - \frac{1}{\operatorname{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell,$$

   puis en déduire une formulation variationnelle du critère $\mathcal{L}(\boldsymbol{I})$ en fonction des lignes $h_i^\top = \frac{u_i^\top}{\sqrt{d_i}}$ de la matrice H.

2. (2 points (bonus)) Proposer une variante de l'algorithme des k-moyennes construisant une suite de partitions $(\boldsymbol{I}_t)_{t\ge 1}$ telle que la suite $(\mathcal{L}(\boldsymbol{I}_t))_{t\ge 1}$ soit décroissante (on justifiera ce point).

Exercice 1 (Questions de cours, 4 points)

1. (1 point) Soient $\{(x_i, y_i)\}_{1\leq i\leq n} \subset \mathbb{R}^d \times \{\pm 1\}$ et $C > 0$. Construire un classifieur SVM revient à déterminer

$$(\hat{w}_n, \hat{b}_n) \in \arg\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|_{\ell_2}^2 + C\sum_{i=1}^{n} \max\left(0, 1 - y_i(w^\top x_i + b)\right).$$

a) Expliquer le rôle de chacun des deux termes dans la fonction à minimiser.

(a) $\hat{w}_n$ est le vecteur de poids, il détermine l'orientation de l'hyperplan.

$\hat{b}_n$ est le terme de biais, il détermine la position de l'hyperplan dans le système de coordonnées.

b) Quelle est la particularité de ce modèle par rapport à celui de régression logistique ?

(b) La régression logistique est un modéle probabiliste. Elle détermine la frontière de classification en ajustant la vraisemblance logarithmique des données et en trouvant l'hyperplan de séparation optimal pour maximiser la probabilité a posteriori des données. Sa frontière de classification est linéaire.

La SVM est un modèle discriminant. La SVM trouver un hyperplan qui maximise la marge de classification. En choisissant une fonction noyau appropriée, les données peuvent être projetées dans

日期: /

un espace de grande dimension où il est possible de trouver un hyperplan linéaire qui les sépare, permettant ainsi de réaliser des classifications non linéaires complexes.

Sa frontière peut-être linéaire et non-linéaire (Kernel Trick).

2. (1 point) Soient $\{(x_i, y_i)\}_{1\leq i\leq n} \subset \mathbb{R}^d \times \mathbb{R}$, $\lambda > 0$ et $\mathcal{H}$ est RKHS de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. On définit alors

$$L : h \in \mathcal{H} \mapsto \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \sum_{i=1}^{n} \|h(x_i) - y_i\|_{\ell_2}^2 .$$

a) Pour tout $h \in \mathcal{H}$, on appelle $h_{\|}$ la projection de h sur $\mathrm{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ et $h_{\perp} = h - h_{\|}$. Montrer que $\sum_{i=1}^{n} \|h(x_i) - y_i\|_{\ell_2}^2 = \sum_{i=1}^{n} \|h_{\|}(x_i) - y_i\|_{\ell_2}^2$.

(a) Comme $\mathcal{H}$ est RKHS de noyau k, on a

$$\forall x \in \mathbb{R}^d,\ \forall h \in \mathcal{H},\quad h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}}$$

et comme $h^{\perp} \in \mathrm{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}^{\perp}$,

$$\begin{aligned} h(x_i) &= \langle h, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle h_{\|} + h^{\perp}, k(\cdot, x_i)\rangle \\ &= \langle h_{\|}, k(\cdot, x_i)\rangle_{\mathcal{H}} + 0 \\ &= h_{\|}(x_i) \end{aligned}$$

donc

$$\sum_{i=1}^{n} \|h(x_i) - y_i\|_{\ell_2}^2 = \sum_{i=1}^{n} \|h_{\|}(x_i) - y_i\|_{\ell_2}^2$$

b) En déduire que $L(h) \geq L(h_{\|})$.

(b) D'après (a), pour déduire que $L(h) \geqslant L(h_{\|})$
il suffit de montrer que $\frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 \geqslant \frac{\lambda}{2}\|h_{\|}\|_{\mathcal{H}}^2$

Et on a

$$\begin{aligned}\|h\|_{\mathcal{H}}^2 &= \langle h, h\rangle_{\mathcal{H}} = \langle h_\perp + h_{\|}, h_\perp + h_{\|}\rangle_{\mathcal{H}} \\ &= \langle h_\perp, h_\perp\rangle_{\mathcal{H}} + \langle h_{\|}, h_{\|}\rangle_{\mathcal{H}} + 2\cdot\langle h_\perp, h_{\|}\rangle_{\mathcal{H}} \\ &= \|h_\perp\|_{\mathcal{H}}^2 + \|h_{\|}\|_{\mathcal{H}}^2 + 0 \\ &\geqslant \|h_{\|}\|_{\mathcal{H}}^2\end{aligned}$$

Donc

$$L(h) \geqslant L(h_{\|})$$

3. (2 points) Soient $r > 0$ et $x \in \mathbb{R}^d$ tel que $\|x\|_{\ell_2} > r$. On souhaite retrouver la projection de x sur la boule de rayon r, $\mathcal{B}_r = \{y \in \mathbb{R}^d, \|y\|_{\ell_2} \leq r\}$, par dualité lagrangienne. Pour ce faire, on résout

$$\begin{array}{ll}\underset{y\in\mathbb{R}^d}{\text{minimiser}} & \|y - x\|_{\ell_2}^2 \\ \text{s. c.} & \|y\|_{\ell_2}^2 \leq r^2.\end{array}$$

a) Vérifier que le problème et convexe et que les conditions de qualification de Slater s'appliquent.

(a) Pour le problème, soit

$$f(y) = \|y - x\|_{\ell_2}^2$$

$$g(y) = \|y\|_{\ell_2}^2 - r^2 \qquad h(y) = 0$$

comme f et g sont fonctions convexes, h est affine
alors le problème est convexe.

Et $\forall y \in \mathbb{R}^d$, $g(y) = \|y\|_{\ell_2}^2 - r^2 \leq 0$

$h(y) = 0$

alors les conditions de qualification de Slater s'appliquent.

b) Définir un lagrangien pour ce problème.

(b) Soit $\lambda \in \mathbb{R}_+$, on peut définir le lagrangien

$$L : (y, \lambda) \in \mathbb{R}^d \times \mathbb{R}_+ \longrightarrow f(y) + \lambda \cdot g(y)$$

$$L(y, \lambda) = \|y - x\|_{\ell_2}^2 + \lambda \cdot (\|y\|_{\ell_2}^2 - r^2)$$

c) Énoncer les conditions KKT.

(c) <u>Les conditions KKT</u>

Si le problème est convexe et si les conditions de Slater sont vérifiées. Pour $x \in \mathbb{R}^d$, $(\lambda, \nu) \in \mathbb{R}^n \times \mathbb{R}^m$.

Alors x^* est la solution du probléme primal et (λ^*, ν^*) est la solution du problème dual, ssi

① $g_i(x^*) \leq 0$ et $h_j(x^*) = 0$, $\forall i, j$

② $\lambda^* \geq 0$

③ $\forall i \in [n]$, $\lambda_i^* = 0$ ou $g_i(x^*) = 0$

④ $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$

d) En déduire une expression de la projection de x sur $\mathcal{B}_r$.

(d) Comme $x \in \mathbb{R}^d$ tel que $\|x\|_{\ell_2} > r$, alors la expression de la projection de x sur B_r est

$$\mathrm{Proj}_{B_r}(x) = \frac{rx}{\|x\|_{\ell_2}}$$

Exercice 2 (Algorithme EM, 7½ points)

Soient $\{(X_i, Y_i)\}_{1\le i\le n}$ des paires indépendantes de variables aléatoires à valeurs dans $\mathbb{R} \times \{b_1, b_2\}$, avec $\{b_1, b_2\} \subset \mathbb{R}$, telles que pour tout $i \in [\![1, n]\!]$:

$$\mathbb{P}(Y_i = b_1) = \alpha_0 \qquad \text{et} \qquad X_i \,|\, Y_i \sim \mathcal{N}\left(a_i^\top \beta_0 + Y_i, \sigma_0^2\right),$$

où $(\alpha_0, \beta_0, \sigma_0^2) \in]0, 1[\times \mathbb{R}^d \times \mathbb{R}_+^*$ est un jeu de paramètres inconnus (les autres sont connus). On suppose ici que l'on n'observe que $\{X_1, \ldots, X_n\}$ et l'on souhaite estimer $(\alpha_0, \beta_0, \sigma_0^2)$ par l'algorithme EM.

1. (1 point) Montrer que la formulation :

$$\begin{cases} X_i = a_i^\top \beta_0 + Y_i + \epsilon_i, \forall i \in [\![1, n]\!] \\ \epsilon \sim \mathcal{N}(0, \sigma_0^2 I_n) \\ \epsilon \perp\!\!\!\perp (Y_1, \ldots, Y_n) \\ \{Y_1, \ldots, Y_n\} \text{ i.i.d avec } \mathbb{P}(Y_1 = b_1) = \alpha_0 \end{cases}$$

est compatible avec le modèle posé (en particulier, on pourra utiliser que deux couples (X_i, Y_i) et (X_j, Y_j), pour $i \neq j$ dans $[\![1, n]\!]$, sont indépendants si pour toutes fonctions boréliennes bornées φ et ψ, $\mathbb{E}[\varphi(X_i, Y_i)\psi(X_j, Y_j)] = \mathbb{E}[\varphi(X_i, Y_i)]\,\mathbb{E}[\psi(X_j, Y_j)]$).

(1) $\Longrightarrow$

Si $\{(X_i,Y_i)\}_{1\le i\le n}$ vérifie la formulation,

alors $\mathbb{P}(Y_i = b_1) = \alpha_0$, $\forall i \in [\![1,n]\!]$

et d'après $\varepsilon \perp\!\!\!\perp (Y_1,\ldots,Y_n)$, alors $\forall i \in [\![1,n]\!]$,

$$E[X_i|Y_i] = a_i^T\beta_0 + Y_i + E[\varepsilon_i|Y_i] = a_i^T\beta_0 + Y_i$$

alors

$$X_i | Y_i \sim \mathcal{N}(a_i^T\beta_0 + Y_i, \sigma_0^2)$$

donc $\{(X_i,Y_i)\}_{1\le i\le n}$ vérifie le modéle posé.

$\Longleftarrow$

Si $\{(X_i,Y_i)\}_{1\le i\le n}$ vérifie le modéle posé,

alors $\{Y_1,\ldots,Y_n\}$ i.i.d. avec $\mathbb{P}(Y_1 = b_1) = \alpha_0$

soit $\varepsilon \perp\!\!\!\perp (Y_1,\ldots,Y_n)$ tel que $\varepsilon \sim \mathcal{N}(0, \sigma_0^2 I_n)$

D'après $X_i | Y_i \sim \mathcal{N}(a_i^T\beta_0 + Y_i, \sigma_0^2)$, $\forall i \in [\![1,n]\!]$,

alors $X_i = a_i^T\beta_0 + Y_i + \varepsilon_i$

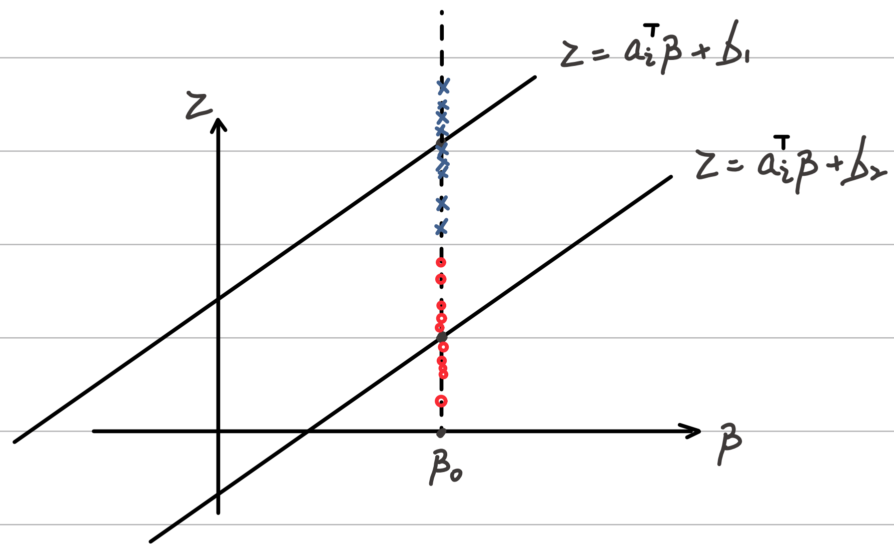
donc $\{(X_i, Y_i)\}_{1\leq i\leq n}$ vérifie la formulation.

Donc la formulation est compatible avec le modéle posé.

2. (1 point) En déduire une interprétation dudit modèle (on pourra se placer dans le cas $d = 1$ et proposer une représentation graphique).
Le modèle correspond-il à un problème de classification ou de régression ? À plan d'expérience fixé (*fixed design*) ou aléatoire ?

(2) L'interprétation

Pour le cas $d=1$, $\forall i \in [\![1,n]\!]$,



Le modéle correspond à un probléme de classification.

À plan d'expérience aléatoire.

3. (1 point) Donner la loi jointe de (X_1, Y_1) (on précisera une mesure dominante). En déduire un modèle statistique pour la loi de (X_1, Y_1) puis l'expression de la log-vraisemblance $\ell_{(X,Y)_1^n}(\alpha, \beta, \sigma^2)$ d'un paramètre quelconque $(\alpha, \beta, \sigma^2) \in]0,1[\times\mathbb{R}^d \times \mathbb{R}_+^*$ (au regard de $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$).

(3)

$$f_{(X_1,Y_1)}(x,y) = \left[\alpha_0 \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot \exp\left(-\frac{(x - a_1^T\beta_0 - b_1)^2}{2\sigma_0^2}\right)\right]^{\mathbb{1}_{\{y=b_1\}}}$$
$$\times \left[\cdot(1-\alpha_0) \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x - a_1^T\beta_0 - b_2)^2}{2\sigma_0^2}\right)\right]^{\mathbb{1}_{\{y=b_2\}}}$$

soit $\delta_{\{b_1\}}$ et $\delta_{\{b_2\}}$ sont mesure de Dirac,

on a

$$(X_1, Y_1) \overset{\mathcal{L}}{\sim} \delta_{\{b_1\}} \cdot \alpha_0 \cdot \mathcal{N}(a_1^T\beta_0 + b_1, \sigma_0^2) + \delta_{\{b_2\}} \cdot (1-\alpha_0) \cdot \mathcal{N}(a_1^T\beta_0 + b_2, \sigma_0^2)$$

<u>Modèle statistique</u>

X 的 modéle statistique

$$\mathcal{P}_m = \left\{ \alpha \cdot \mathcal{N}(a_i^T\beta + b_1, \sigma^2) + (1-\alpha) \cdot \mathcal{N}(a_i^T\beta + b_2, \sigma^2) : \theta = (\alpha, \beta, \sigma^2) \in \Theta \right\}$$

<u>log-vraisemblance</u>

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

$$\log f_{(X_i,Y_i)}(\alpha, \beta, \sigma^2) = \mathbb{1}_{\{Y_i=b_1\}} \cdot \left(\log\alpha - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_i - a_i^T\beta - b_1)^2}{2\sigma^2}\right)$$
$$+ \mathbb{1}_{\{Y_i=b_2\}} \cdot \left(\log(1-\alpha) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_i - a_i^T\beta - b_2)^2}{2\sigma^2}\right)$$

alors

$$\ell_{(X,Y)_1^n}(\alpha, \beta, \sigma^2) = \sum_{i=1}^n \mathbb{1}_{\{Y_i=b_1\}} \cdot \left(\log\alpha - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_i - a_i^T\beta - b_1)^2}{2\sigma^2}\right)$$
$$+ \sum_{i=1}^n \mathbb{1}_{\{Y_i=b_2\}} \cdot \left(\log(1-\alpha) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_i - a_i^T\beta - b_2)^2}{2\sigma^2}\right)$$

4. (1 point) Expliciter, pour tout $x \in \mathbb{R}$, la loi de $Y_1 \mid X_1 = x$.

On suppose disposer d'un estimateur candidat $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ de $(\alpha_0, \beta_0, \sigma_0^2)$. Construire n variables aléatoires $Z_1, \ldots, Z_n$ visant à « approcher » $Y_1, \ldots, Y_n$, connaissant l'estimateur $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$.

(4) La loi de $Y_1 | X_1 = x$

D'après (3), on a

$$\mathbb{P}(Y_1 = b_1 | X = x) = \alpha_0 \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot \exp\left(-\frac{(x - a_1^\top \beta_0 - b_1)^2}{2\sigma_0^2}\right)$$

$$\mathbb{P}(Y_1 = b_2 | X = x) = (1 - \alpha_0) \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot \exp\left(-\frac{(x - a_1^\top \beta_0 - b_1)^2}{2\sigma_0^2}\right)$$

On suppose disposer d'un estimateur candidat $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ de $(\alpha_0, \beta_0, \sigma_0^2)$. Construire n variables aléatoires $Z_1, \ldots, Z_n$ visant à « approcher » $Y_1, \ldots, Y_n$, connaissant l'estimateur $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$.

5. (1 point) En déduire que pour tout $(\alpha, \beta, \sigma^2) \in]0,1[\times \mathbb{R}^d \times \mathbb{R}_+^*$,

$$\begin{aligned}
&\mathbb{E}[\ell_{(X,Z)_1^n}(\alpha, \beta, \sigma^2) \mid X_1, \ldots, X_n] \\
&= \log(\alpha) \sum_{i=1}^n p_i + \log(1-\alpha)\left(n - \sum_{i=1}^n p_i\right) - \frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) \\
&- \frac{1}{2\sigma^2} \sum_{i=1}^n \left[p_i \left(X_i - (a_i^\top \beta + b_1)\right)^2 + (1 - p_i)\left(X_i - (a_i^\top \beta + b_2)\right)^2\right],
\end{aligned}$$

où $p_1, \ldots, p_n$ sont à déterminer.

(5) $\mathbb{E}[\ell_{(X,Z)_1^n}(\alpha, \beta, \sigma^2) | X_1, \ldots, X_n]$

$$= \sum_{i=1}^n \mathbb{P}(Z_i = b_1 | X_i) \cdot \left(\log(\alpha) - \frac{1}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{(X_i - (a_i^\top \beta + b_1))^2}{2\sigma^2}\right)$$

$$+ \sum_{i=1}^n (1 - \mathbb{P}(Z_i = b_1 | X_i)) \cdot \left(\log(1-\alpha) - \frac{1}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{(X_i - (a_i^\top \beta + b_2))^2}{2\sigma^2}\right)$$

$$= \log(\alpha) \cdot \sum_{i=1}^n \mathbb{P}(Z_i = b_1 | X_i) + \log(1-\alpha) \cdot \left(n - \sum_{i=1}^n p_i\right) - \frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^n \left[\mathbb{P}(Z_i = b_1 | X_i) \cdot (X_i - (a_i^\top \beta + b_1))^2 + (1 - \mathbb{P}(Z_i = b_1 | X_i)) \cdot (X_i - (a_i^\top \beta + b_2))^2\right]$$

$\forall i \in [\![1, n]\!]$, on a

$$p_i = \mathbb{P}(Z_i = b_1 | X_i) = \hat{\alpha} \cdot \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \cdot \exp\left(-\frac{(X_i - a_i^\top \hat{\beta} - b_1)^2}{2\hat{\sigma}^2}\right)$$

6. (2 points) En appelant $A = \begin{pmatrix} a_1^\top \\ \vdots \\ a_n^\top \end{pmatrix} \in \mathbb{R}^{n\times d}$ et en supposant que rang$(A) = d$, déterminer

$$\arg\max_{(\alpha,\beta,\sigma^2)\in]0,1[\times\mathbb{R}^d\times\mathbb{R}_+^*} \mathbb{E}[\ell_{(X,Z)_1^n}(\alpha,\beta,\sigma^2) \mid X_1,\dots,X_n].$$

(6) soit $F(\alpha,\beta,\sigma^2) = \mathbb{E}\left[\ell_{(X,Z)_1^n}(\alpha,\beta,\sigma^2) \middle| X_1,\dots,X_n\right]$

① $\nabla_\alpha F(\alpha,\beta,\sigma^2) = \frac{1}{\alpha}\cdot\sum_{i=1}^n p_i - \frac{1}{1-\alpha}\cdot\left(n-\sum_{i=1}^n p_i\right)$

on prend $\nabla_\alpha F(\alpha,\beta,\sigma^2) = 0$

$$\hat\alpha = \frac{1}{n}\sum_{i=1}^n p_i$$

② $\nabla_\beta F(\alpha,\beta,\sigma^2)$

$$= -\frac{1}{2\sigma^2}\cdot\sum_{i=1}^n\left[-2\cdot a_i^\top\cdot p_i\left(X_i-(a_i^\top\beta+b_1)\right) - 2\cdot a_i^\top\cdot(1-p_i)\cdot\left(X_i-(a_i^\top\beta+b_2)\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^n a_i^\top\cdot\left(X_i - a_i^\top\beta - p_i\cdot b_1 - (1-p_i)\cdot b_2\right)$$

on prend $\nabla_\beta F(\alpha,\beta,\sigma^2) = 0$

alors $a_i^\top\hat\beta = X_i - p_i b_1 - (1-p_i) b_2$, $\forall i \in [\![1,n]\!]$

donc
$$A\cdot\hat\beta = \begin{bmatrix} X_1 - p_1 b_1 - (1-p_1) b_2 \\ \vdots \\ X_n - p_n b_1 - (1-p_n) b_2 \end{bmatrix}$$

$$\hat\beta = A^{-1}\begin{bmatrix} X_1 - p_1 b_1 - (1-p_1) b_2 \\ \vdots \\ X_n - p_n b_1 - (1-p_n) b_2 \end{bmatrix}$$

③ $\nabla_{\sigma^2} F(\alpha,\beta,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\cdot\sum_{i=1}^n\left[p_i\left(X_i-(a_i^\top\beta+b_1)\right)^2 + (1-p_i)\left(X_i-(a_i^\top\beta+b_2)\right)^2\right]$

on prend $\nabla_{\sigma^2} F(\alpha,\beta,\sigma^2) = 0$

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n\left[p_i\left(X_i-(a_i^\top\beta+b_1)\right)^2 + (1-p_i)\left(X_i-(a_i^\top\beta+b_2)\right)^2\right]$$

7. (½ point) Décrire l'algorithme EM adapté au problème posé.

(7) Pour l'algorithme EM

initialiser $(\hat{\alpha}^{(0)}, \hat{\beta}^{(0)}, \hat{\sigma}^{2(0)}) = \theta^{(0)}$

pour le $t$ème itération :

① (E-step) Calculer

$$p_i^{(t+1)} = \mathbb{P}(Z_i = b_1 \mid X_i, \theta^{(t+1)})$$

② (M-step) Calculer

$$\hat{\alpha}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_i^{(t)} \qquad \hat{\beta}^{(t+1)} = A^{-1}U^{(t)}$$

$$\hat{\sigma}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[p_i^{(t)}(X_i-(a_i^T\hat{\beta}^{(t)}+b_1))^2+(1-p_i^{(t)})(X_i-(a_i^T\hat{\beta}^{(t)}+b_2))^2\right]$$

伪代码格式 EM algorithm for Gaussien Mixtures

Input : $\{X_i\}_{1\leq i\leq n}$

$p_i \longleftarrow \frac{1}{n}$ for all $i \in [\![1,n]\!]$

$\alpha \longleftarrow \frac{1}{2}$

$\beta \longleftarrow$ random initialisation

$\sigma^2 \longleftarrow$ random initialisation

while not converged do :

$p_i \longleftarrow \alpha \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(X_i - a_i^T\beta - b_1)^2}{2\sigma^2}\right)$, $\forall i \in [\![1,n]\!]$

$\alpha \longleftarrow \frac{1}{n}\sum_{i=1}^{n} p_i$

$\beta \longleftarrow A^{-1}U$

$\sigma^2 \longleftarrow \frac{1}{n}\sum_{i=1}^{n}\left[p_i(X_i-(a_i^T\beta+b_1))^2+(1-p_i)(X_i-(a_i^T\beta+b_2))^2\right]$

end while

Exercice 3 (Clustering spectral, 8½ points)

Soient $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ et $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ une mesure de similarité symétrique. On appelle $W = (s(x_i, x_j))_{1 \leq i,j \leq n}$ la matrice d'adjacence des données, $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$ la matrice diagonale des degrés $d_i = \sum_{j=1}^n W_{ij}$ et $L = D - W$ le laplacien non-normalisé du graphe associé. Soient de plus $L_s = D^{-1/2} L D^{-1/2}$ et $L_w = D^{-1} L$ les laplaciens normalisés.

Préliminaires

非奇异 $\longleftrightarrow$ 可逆

1. (½ point) Donner une condition suffisante sur s pour que D soit non-singulière.

(1) Pour que D soit non-singulière,

$$\begin{cases} \forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d, & s(x, x') = s(x', x) \\ \forall x \in \mathbb{R}^d, & s(x, x) > 0 \end{cases}$$

2. (½ point) On suppose D non-singulière. Montrer que $u \in \mathbb{R}^n$ est vecteur propre de L_w avec pour valeur propre $\lambda \in \mathbb{R}$ si est seulement si $D^{1/2}u$ est vecteur propre de L_s avec pour valeur propre λ.

(2) $u \in \mathbb{R}^n$ est vecteur propre de L_w avec pour valeur propre $\lambda \in \mathbb{R}$

$\Longleftrightarrow$

$$L_w u = \lambda u$$

$$D^{-1} L u = \lambda u$$

$$D^{-\frac{1}{2}} D^{-\frac{1}{2}} L D^{-\frac{1}{2}} D^{\frac{1}{2}} u = \lambda u$$

$$L_s D^{\frac{1}{2}} u = \lambda D^{\frac{1}{2}} u$$

$\Longleftrightarrow$

$D^{\frac{1}{2}} u$ est vecteur propre de L_s avec pour valeur propre $\lambda \in \mathbb{R}$.

3. (½ point) En déduire que $\mathbb{1}$ et $(\sqrt{d_1}, \ldots, \sqrt{d_n})$ sont vecteurs propres de L_w et L_s respectivement et déterminer les valeurs propres associées.

(3) ① Pour $\mathbb{1} = (1, \ldots, 1)^T$

$$L_w \mathbb{1} = D^{-1} L \mathbb{1}$$

$$= \begin{bmatrix} \frac{1}{d_1} & & \\ & \ddots & \\ & & \frac{1}{d_n} \end{bmatrix} \cdot \begin{bmatrix} d_1 - w_{11} & -w_{12} & \cdots & -w_{1n} \\ -w_{21} & d_2 - w_{22} & \cdots & -w_{2n} \\ \vdots & \vdots & & \vdots \\ -w_{n1} & -w_{n2} & \cdots & d_n - w_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= D^{-1} \cdot 0$$

$$= 0 \cdot \mathbb{1}$$

donc la valeur propre pour L_w est 0

② Pour $(\sqrt{d_1}, \ldots, \sqrt{d_n}) = D^{\frac{1}{2}} \cdot \mathbb{1}$

d'après (2), la valeur propre pour L_s est 0

Partie A

Pour une partie $I \subset [\![1, n]\!]$, on définit $\mathrm{vol}(I) = \sum_{i \in I} d_i$ et le vecteur

$$f_I = \left(\sqrt{\frac{\mathrm{vol}(I^c)}{\mathrm{vol}(I)}} \mathbf{1}_{i \in I} - \sqrt{\frac{\mathrm{vol}(I)}{\mathrm{vol}(I^c)}} \mathbf{1}_{i \in I^c} \right)_{1 \leq i \leq n},$$

où I^c est le complémentaire de I dans $[\![1, n]\!]$.

1. (1 point) Montrer que pour $i \in I$ et $j \in I^c$, en notant f_{Ii} la i^e composante de f_I,

$$(f_{Ii} - f_{Ij})^2 = \frac{\mathrm{tr}(D)}{\mathrm{vol}(I)} + \frac{\mathrm{tr}(D)}{\mathrm{vol}(I^c)}.$$

(1) Pour $i \in I$ et $j \in I^c$,

$$f_{Ii} = \sqrt{\frac{\mathrm{vol}(I^c)}{\mathrm{vol}(I)}} \qquad f_{Ij} = -\sqrt{\frac{\mathrm{vol}(I)}{\mathrm{vol}(I^c)}}$$

alors

$$(f_{Ii} - f_{Ij})^2 = f_{Ii}^2 + f_{Ij}^2 + 2 \cdot f_{Ii} \cdot f_{Ij}$$

$$= \frac{\mathrm{vol}(I^c)}{\mathrm{vol}(I)} + \frac{\mathrm{vol}(I)}{\mathrm{vol}(I^c)} + 2$$

$$= \frac{\mathrm{vol}(I^c) + \mathrm{vol}(I)}{\mathrm{vol}(I)} + \frac{\mathrm{vol}(I) + \mathrm{vol}(I^c)}{\mathrm{vol}(I^c)}$$

$$= \frac{\mathrm{tr}(D)}{\mathrm{vol}(I)} + \frac{\mathrm{tr}(D)}{\mathrm{vol}(I^c)}$$

2. (1 point) Montrer que $\mathbb{1}^\top D f_I = 0$ et $f_I^\top D f_I = \mathrm{tr}(D)$.

(2)

$$\mathbb{1}^T D f_I = [d_1, \dots, d_n] f_I$$
$$= \sum_{i=1}^{n} d_i \times f_{I_i}$$
$$= \sqrt{\frac{vol(I^c)}{vol(I)}} \cdot \sum_{i \in I} d_i - \sqrt{\frac{vol(I)}{vol(I^c)}} \cdot \sum_{j \in I^c} d_j$$
$$= \sqrt{vol(I) \cdot vol(I^c)} - \sqrt{vol(I) \cdot vol(I^c)}$$
$$= 0$$

$$f_I^T D f_I = \sum_{i=1}^{n} d_i \cdot (f_{I_i})^2$$
$$= \frac{vol(I^c)}{vol(I)} \cdot \sum_{i \in I} d_i + \frac{vol(I)}{vol(I^c)} \cdot \sum_{j \in I^c} d_j$$
$$= vol(I^c) + vol(I)$$
$$= tr(D)$$

3. (1 point) Sachant que pour tout $u \in \mathbb{R}^n, u^\top L u = \frac{1}{2}\sum_{1\leq i,j\leq n} W_{ij}(u_i - u_j)^2$, montrer que

$$f_I^\top L f_I = \mathrm{tr}(D)\left(\frac{\sum_{\substack{i\in I\\ j\in I^c}} W_{ij}}{\mathrm{vol}(I)} + \frac{\sum_{\substack{i\in I^c\\ j\in I}} W_{ij}}{\mathrm{vol}(I^c)}\right).$$

(3)

$$f_I^\top L f_I = \frac{1}{2} \cdot \sum_{1\leq i,j\leq n} W_{i,j} \cdot (f_{I_i} - f_{I_j})^2$$

$$= \frac{1}{2} \cdot \left(0 + \sum_{\substack{i\in I\\ j\in I^c}} W_{i,j} \cdot (f_{I_i} - f_{I_j})^2 + \sum_{\substack{i\in I^c\\ j\in I}} W_{i,j} \cdot (f_{I_i} - f_{I_j})^2 + 0 \right)$$

$$= \frac{1}{2} \left(\sum_{\substack{i\in I\\ j\in I^c}} W_{i,j} \cdot \left(\frac{tr(D)}{vol(I)} + \frac{tr(D)}{vol(I^c)} \right) + \sum_{\substack{i\in I^c\\ j\in I}} W_{i,j} \cdot \left(\frac{tr(D)}{vol(I)} + \frac{tr(D)}{vol(I^c)} \right) \right)$$

$$= tr(D) \cdot \left(\frac{\sum_{\substack{i\in I\\ j\in I^c}} W_{ij}}{vol(I)} + \frac{\sum_{\substack{i\in I^c\\ j\in I}} W_{ij}}{vol(I^c)} \right)$$

4. (1 point) En déduire une réécriture du problème de *Normalized cut* dans le cas d'une seule coupure (*i.e.* d'un partitionnement en deux groupes) et un relâchement de celui-ci sous la forme d'un problème d'optimisation continue.

(4) Pour le problème de Normalized cut,

$$\underset{I, I^c \subset P([n])}{\text{minimize}} \quad \frac{\sum_{\substack{i\in I\\ j\in I^c}} W_{ij}}{Vol(I)} + \frac{\sum_{\substack{i\in I^c\\ j\in I}} W_{ij}}{vol(I^c)}$$

d'après (3), il est équivalent à

$$\underset{I, I^c \subset P([n])}{\text{minimize}} \quad f_I^\top L f_I$$

et le relâchement de ce probléme est

日期: /

$$\underset{F \subset \mathbb{R}^n}{\text{minimize}} \; F^T L F$$

$$\text{s.c} \begin{cases} \mathbb{1}^T D F = 0 \\ F^T D F = tr(D) \end{cases}$$

5. (1 point) Expliciter une solution du problème relâché.

Partie B

On s'intéresse à présent au problème de partitionnement en k (entier supérieur à 2) groupes par clustering spectral minimisant le coût *Normalized cut*, et on nomme donc $U \in \mathbb{R}^{n\times k}$ la matrice dont les colonnes sont les vecteurs propres de L_s associés aux k plus petites valeurs propres. On souhaite, à partir de cette matrice, remonter à une partition $\boldsymbol{I} = (I_1, \ldots, I_k)$ de $[\![1, n]\!]$ telle que « le sous-espace vectoriel engendré par la partition $\boldsymbol{I}$ » soit aussi proche que possible de celui engendré par les colonnes de $H = D^{-1/2}U$, la matrice solution du problème relâché. Autrement dit, on souhaite avoir $\text{range}(D^{1/2}Y_{\boldsymbol{I}}) \approx \text{range}(U)$, où $Y_{\boldsymbol{I}} = (\mathbf{1}_{i\in I_j})_{\substack{1\le i\le n\\ 1\le j\le k}} \in \mathbb{R}^{n\times k}$ est la matrice *one-hot encoding* de la partition $\boldsymbol{I}$.

Pour ce faire, on utilise une distance entre projecteurs :

$$\mathcal{L}(\boldsymbol{I}) = \frac{1}{2}\|P_U - P_{\boldsymbol{I}}\|_F^2 = k - \sum_{j=1}^{k} \frac{1}{\text{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell,$$

où P_U et $P_{\boldsymbol{I}}$ sont les projecteurs orthogonaux sur $\text{range}(U)$ et $\text{range}(D^{1/2}Y_{\boldsymbol{I}})$ respectivement, $u_i^\top$ est la i^e ligne de U et où l'on a remarqué que $\sum_{\substack{1\le i\le n\\ 1\le j\le k}} U_{ij}^2 = k$.

1. (2 points) Montrer que, pour tout $j \in [\![1, k]\!]$,

$$\min_{\mu\in\mathbb{R}^k} \sum_{i\in I_j} d_i \left\| \frac{u_i}{\sqrt{d_i}} - \mu \right\|_{\ell_2}^2 = \sum_{i\in I_j} \|u_i\|_{\ell_2}^2 - \frac{1}{\text{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell,$$

puis en déduire une formulation variationnelle du critère $\mathcal{L}(\boldsymbol{I})$ en fonction des lignes $h_i^\top = \frac{u_i^\top}{\sqrt{d_i}}$ de la matrice H.

(1) $$\sum_{i\in I_j} d_i \cdot \left\| \frac{u_i}{\sqrt{d_i}} - \mu \right\|_{\ell_2}^2 = \sum_{i\in I_j} \| u_i - \mu\sqrt{d_i} \|_{\ell_2}^2$$

$$= \sum_{i\in I_j} \|u_i\|_{\ell_2}^2 + \|\mu\sqrt{d_i}\|_{\ell_2}^2 - 2\langle u_i, \mu\sqrt{d_i}\rangle_{\ell_2}$$

$$= \sum_{i\in I_j} \|u_i\|_{\ell_2}^2 + \sum_{i\in I_j}\left[d_i \cdot \|\mu\|_{\ell_2}^2 - 2\sqrt{d_i}\cdot u_i^\top \mu\right]$$

D'après la propriété de la fonction quadratique.

...

$$\min_{\mu\in\mathbb{R}^k} \sum_{i\in I_j}\left[d_i \cdot \|\mu\|_{\ell_2}^2 - 2\sqrt{d_i}\cdot u_i^\top \mu\right] = -\frac{1}{\text{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell$$

donc

$$\min_{\mu\in\mathbb{R}^k} \sum_{i\in I_j} d_i \cdot \left\| \frac{u_i}{\sqrt{d_i}} - \mu \right\|_{\ell_2}^2 = \sum_{i\in I_j} \|u_i\|_{\ell_2}^2 - \frac{1}{\text{vol}(I_j)} \sum_{i,\ell\in I_j} \sqrt{d_i d_\ell}\, u_i^\top u_\ell$$

日期: /

$$\mathcal{L}(\mathcal{I}) = k - \sum_{j=1}^{k} \frac{1}{\mathrm{Vol}(\mathcal{I}_j)} \cdot \sum_{i,l \in \mathcal{I}_j} \sqrt{d_i d_l}\, u_i^T u_l$$

$$= k + \sum_{j=1}^{k} \left(\min_{\mu \in \mathbb{R}^k} \sum_{i \in \mathcal{I}_j} d_i \cdot \|h_i - \mu\|_{\ell_2}^2 - \sum_{i \in \mathcal{I}_j} \|u_i\|_{\ell_2}^2 \right)$$

$$= \sum_{j=1}^{k} \left(\min_{\mu \in \mathbb{R}^k} \sum_{i \in \mathcal{I}_j} d_i \cdot \|h_i - \mu\|_{\ell_2}^2 + 1 - \sum_{i \in \mathcal{I}_j} \|u_i\|_{\ell_2}^2 \right)$$

$$= \min_{\mu \in \mathbb{R}^k} \sum_{j=1}^{k} \left(\sum_{i \in \mathcal{I}_j} d_i \cdot \|h_i - \mu\|_{\ell_2}^2 + 1 - \sum_{i \in \mathcal{I}_j} d_i \|h_i\|_{\ell_2}^2 \right)$$

2. (2 points (bonus)) Proposer une variante de l'algorithme des k-moyennes construisant une suite de partitions $(\boldsymbol{I}_t)_{t\geq 1}$ telle que la suite $(\mathcal{L}(\boldsymbol{I}_t))_{t\geq 1}$ soit décroissante (on justifiera ce point).

(2) D'après (1), on a un probléme d'optimisation

$$\underset{\substack{I_1,\dots,I_k \in \mathcal{P}([k]) \\ \mu \in \mathbb{R}^k}}{\text{minimise}} \sum_{j=1}^{k} \left(\sum_{i \in I_j} d_i \cdot \|h_i - \mu\|_{\ell_2}^2 + 1 - \sum_{i \in I_j} d_i \|h_i\|_{\ell_2}^2 \right)$$

la variante de l'algorithm des k-moyenne

Input : $T \in \mathbb{N}$, $\{X_i\}_{1\leq i\leq n}$

$\mu \longleftarrow$ random point from $\mathbb{R}^k$

for $t=1$ to T do

compute a Voronoi partitioning $(I_1,\dots,I_k)$ corresponing μ

$\mu \longleftarrow$

end for

Output : $(I_1,\dots,I_k)$

日期:　　　/