Examen : Introduction à l'apprentissage automatique

18 décembre 2020

Aucun document n'est autorisé.

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

Le barème (sur 20 points, auxquels s'ajoutent 4 points bonus) n'est donné qu'à titre indicatif.

Exercice 1 (Questions de cours, 4 points)

Soient $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $K > 1$ un entier fixé et $\mathcal{P}$ l'ensemble des partitions de $\mathbb{R}^d$ à K cellules.

1. (1 point) On considère le problème d'optimisation

$$\begin{aligned} \underset{\substack{(C_1,\ldots,C_K)\in\mathcal{P} \\ \mu_1,\ldots,\mu_K\in\mathbb{R}^d}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C_j} \\ \text{s. t.} \quad & \mu_j \in \arg\min_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} \|x_i - \mu\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C_j}, \quad \forall j \in [\![1, K]\!]. \end{aligned}$$

   À quoi correspond ce problème et à quel domaine de l'apprentissage automatique appartient-il ? Décrire l'algorithme de résolution associé.

2. (1 point) Soient $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ fixés, $(C_1, \ldots, C_K) \in \mathcal{P}$ le partitionnement de Voronoi associé et $(C'_1, \ldots, C'_K) \in \mathcal{P}$ une partition quelconque. Montrer qu'alors

$$\sum_{i=1}^{n}\sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C_j} \leq \sum_{i=1}^{n}\sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C'_j}.$$

3. (1 point) Soient $\{p_i\}_{1\leq i\leq n} \subset \mathbb{R}_+^*$ fixés. Expliciter :

$$\arg\min_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} p_i \|x_i - \mu\|_{\ell_2}^2 .$$

4. (1 point) Quelles sont les différences entre le problème considéré ici et la méthode appelée *soft K-means* ?

Exercice 2 (Modèle de classification, 6 points)

Rappels sur les lois de probabilité

- R est une variable aléatoire de Rademacher de paramètre $p \in]0,1[$ (on écrit $R \sim \mathcal{R}(p)$) si R est à valeurs dans $\{\pm 1\}$ telle que $\mathbb{P}(R=1)=p$.
- E est une variable aléatoire exponentielle (on écrit $E \sim \mathcal{E}$) si E est à valeurs dans $\mathbb{R}_+$ et de fonction de répartition $x \in \mathbb{R} \mapsto (1-e^{-x})\mathbf{1}_{\mathbb{R}_+}(x)$.

Soient (X,Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$, $f^\star : \mathbb{R}^d \to \mathbb{R}$ une fonction mesurable et ε une variable aléatoire réelle telle que $\varepsilon \,\big|\, X \sim \mathcal{E}$. On considère le modèle de classification :

$$Y = \operatorname{sign}\left(f^\star(X) + \varepsilon\right).$$

1. (1 point) Montrer que $Y|X \sim \mathcal{R}(p_X)$, avec p_X à déterminer.
2. (1 point) Soit $x \in \mathbb{R}^d$. Exprimer $-\log\left(\mathbb{P}(Y=1 \,\big|\, X=x)\right)$ comme fonction de la variable $f^\star(x)$:
$$\ell_1(f^\star(x)) = -\log\left(\mathbb{P}(Y=1 \,\big|\, X=x)\right).$$
Représenter le graphe de ℓ_1. Est-ce une fonction convexe ?
3. (1 point) Même question pour $\ell_{-1}(f^\star(x)) = -\log\left(\mathbb{P}(Y=-1 \,\big|\, X=x)\right)$.
4. (2 points) On considère à présent le modèle linéaire
$$\forall x \in \mathbb{R}^d : \qquad f^\star(x) = {\beta^\star}^\top x,$$
avec $\beta^\star \in \mathbb{R}^d$ inconnu. Comment obtenir un estimateur du maximum de vraisemblance de $\beta^\star$ à partir d'un échantillon $(X_1,Y_1),\dots,(X_n,Y_n)$ de même loi que (X,Y) (expliciter le problème de minimisation associé à la log-vraisemblance avec ses éventuelles contraintes) ?
5. (1 point) Déterminer le gradient de la fonction objectif introduite à la question précédente lorsqu'il existe.

Exercice 3 (Approximation de noyau, 10 points)

Rappels généraux et de concentration

- Vecteur gaussien : un vecteur aléatoire gaussien X de dimension d, de moyenne $\mu \in \mathbb{R}^d$ et de matrice de covariance Σ symétrique définie positive (on notera $X \sim \mathcal{N}(\mu,\Sigma)$) admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, s'exprimant :
$$x \in \mathbb{R}^d \mapsto \frac{1}{\sqrt{2\pi}^d |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$
- Trigonométrie :
$$\forall a,b \in \mathbb{R}, \quad \cos a \cos b = \frac{\cos(a+b)+\cos(a-b)}{2}.$$

— Inégalité de Markov : soit Z une variable aléatoire réelle positive presque sûrement. Alors
$$\forall t > 0, \quad \mathbb{P}\left(Z \geq t\right) \leq \frac{\mathbb{E}[Z^2]}{t^2}.$$

— Inégalité de Hoeffding : soit $(X_N)_{N\geq 1}$ une suite de variables aléatoires réelles indépendantes et bornées par $a > 0$, *i.e.* $|X_i| \leq a$ p.s. pour tout $i \in \mathbb{N}^*$. Alors en notant $S_N = \sum_{i=1}^N X_i$,
$$\forall t > 0, \quad \mathbb{P}\left(|S_N - \mathbb{E}[S_N]| \geq t\right) \leq 2e^{-\frac{t^2}{2Na^2}}.$$

Partie A

1. (1½ points) Soient $\gamma > 0$ et $\omega \sim \mathcal{N}(0, 2\gamma I_d)$. Montrer que pour tout $u \in \mathbb{R}^d$:
$$e^{\gamma\|u\|_{\ell_2}^2}\,\mathbb{E}[e^{i\omega^\top u}] = 1.$$
En déduire que
$$e^{-\gamma\|u\|_{\ell_2}^2} = \mathbb{E}[\cos(\omega^\top u)].$$

2. (1 point) Soit $\varphi \sim \mathcal{U}([0,\pi])$ indépendante de ω. Montrer, par intégration, que pour tout $x, y \in \mathbb{R}^d$:
$$\mathbb{E}[\cos(\omega^\top(x-y))] = 2\,\mathbb{E}[\cos(\omega^\top x + \varphi)\cos(\omega^\top y + \varphi)].$$

3. (1½ points) Soient $(\omega_N)_{N\geq 1}$ une suite de vecteurs aléatoires i.i.d. de même loi que ω et indépendants de $(\varphi_N)_{N\geq 1}$ i.i.d. de même loi que φ :
$$(\omega_N)_{N\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, 2\gamma I_d) \perp\!\!\!\perp (\varphi_N)_{N\geq 1} \overset{i.i.d.}{\sim} \mathcal{U}([0,\pi]).$$
Pour un entier $N > 1$ fixé, et $x, y \in \mathbb{R}^d$ quelconques, on s'intéresse à l'approximation :
$$e^{-\gamma\|x-y\|_{\ell_2}^2} \approx \frac{2}{N}\sum_{i=1}^N \cos(\omega_i^\top x + \varphi_i)\cos(\omega_i^\top y + \varphi_i).$$
Quel est l'intérêt de celle-ci pour l'apprentissage automatique (penser aux aspects algorithmiques lors des phases d'apprentisage et de prédiction) ?

4. (1 point) Montrer que pour tout $\epsilon > 0$:
$$\mathbb{P}\left(\left|\frac{2}{N}\sum_{i=1}^N \cos(\omega_i^\top x + \varphi_i)\cos(\omega_i^\top y + \varphi_i) - e^{-\gamma\|x-y\|_{\ell_2}^2}\right| \geq \epsilon\right) \leq 2e^{-\frac{N\epsilon^2}{8}}.$$

5. (2 points) Soient $\mathcal{S} \subset \mathbb{R}^d$ un ensemble de cardinal $n > 1$ fixé et $\delta \in]0,1[$. Montrer qu'avec probabilité au moins $1-\delta$:
$$\forall x, y \in \mathcal{S}, \quad \left|\frac{2}{N}\sum_{i=1}^N \cos(\omega_i^\top x + \varphi_i)\cos(\omega_i^\top y + \varphi_i) - e^{-\gamma\|x-y\|_{\ell_2}^2}\right| \leq \sqrt{\frac{16}{N}\log\left(\frac{n+1}{\sqrt{\delta}}\right)}.$$

Partie B

On souhaite à présent établir une borne similaire à celle de la question précédente mais pour un ensemble non-dénombrable.

Soient $\mathcal{X} \subset \mathbb{R}^d$ un espace borné par $\frac{1}{8}$: $\sup_{x\in\mathcal{X}} \|x\|_{\ell_2} \leq \frac{1}{8}$; et $r > 0$ un rayon quelconque. On admet qu'il existe $T \leq r^{-d}$ boules de rayons r recouvrant $\mathcal{U} = \{x - y : x, y \in \mathcal{X}\}$ et on note $U_1, \dots, U_T$ leurs centres.

On appelle f_N l'erreur d'approximation, définie pour tout $u \in \mathcal{U}$ par :

$$f_N(u) = \frac{2}{N} \sum_{i=1}^{N} \cos(\omega_i^\top x_u + \varphi_i) \cos(\omega_i^\top y_u + \varphi_i) - e^{-\gamma\|u\|_{\ell_2}^2},$$

où $x_u, y_u \in \mathcal{X}$ sont tels que $u = x_u - y_u$, et on admet que f_N est L_N-lipschitzienne ($L_N > 0,$ p.s.) avec $\sigma^2 = \mathbb{E}[L_N^2] \leq \mathbb{E}[\|\omega\|_{\ell_2}^2]$.

1. (1 point) Montrer que $\sigma^2 \leq \mathrm{tr}(\mathbb{V}(\omega)) = 2d\gamma$.
2. (1 point) Soient $u \in \mathcal{U}$, $i \in [\![1, T]\!]$ tel que $\|u - U_i\|_{\ell_2} \leq r$ et $\epsilon > 0$. Montrer que :
$$\left[|f_N(U_i)| \leq \frac{\epsilon}{2} \text{ et } L_N \leq \frac{\epsilon}{2r}\right] \implies [|f_N(u)| \leq \epsilon] \,.$$
3. (1 point) En déduire que :
$$\mathbb{P}\left(\sup_{u\in\mathcal{U}} |f_N(u)| > \epsilon\right) \leq \sum_{i=1}^{T} \mathbb{P}\left(|f_N(U_i)| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(L_N > \frac{\epsilon}{2r}\right).$$
4. (2 points (bonus))
$$\mathbb{P}\left(\sup_{u\in\mathcal{U}} |f_N(u)| > \epsilon\right) \leq \frac{4r^2\sigma^2}{\epsilon^2} + \frac{2}{r^d} e^{-\frac{N\epsilon^2}{32}}.$$
5. (2 points (bonus)) En remarquant que pour tout $a, b > 0$,
$$ar^2 + \frac{b}{r^d} = 2a^{\frac{d}{d+2}} b^{\frac{2}{d+2}},$$
lorsque $r = \left(\frac{b}{a}\right)^{\frac{1}{d+2}}$, montrer que pour $\epsilon \in]0, \sigma]$:
$$\mathbb{P}\left(\sup_{x,y\in\mathcal{X}} \left| \frac{2}{N} \sum_{i=1}^{N} \cos(\omega_i^\top x + \varphi_i) \cos(\omega_i^\top y + \varphi_i) - e^{-\gamma\|x-y\|_{\ell_2}^2} \right| > \epsilon\right) \leq 2^4 d \frac{\gamma}{\epsilon^2} e^{-\frac{N\epsilon^2}{16(d+2)}}.$$

Exercice 1 (Questions de cours, 4 points)

Soient $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $K > 1$ un entier fixé et $\mathcal{P}$ l'ensemble des partitions de $\mathbb{R}^d$ à K cellules.

1. (1 point) On considère le problème d'optimisation

$$\begin{aligned} \underset{\substack{(C_1,\ldots,C_K)\in\mathcal{P} \\ \mu_1,\ldots,\mu_K\in\mathbb{R}^d}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C_j} \\ \text{s. t.} \quad & \mu_j \in \arg\min_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} \|x_i - \mu\|_{\ell_2}^2 \, \mathbf{1}_{x_i\in C_j}, \quad \forall j \in [\![1, K]\!]. \end{aligned}$$

À quoi correspond ce problème et à quel domaine de l'apprentissage automatique appartient-il ? Décrire l'algorithme de résolution associé.

(1) Ce problème correspond à k-moyenne algorithme.

L'apprentissage automatique appartient à partitionner par minimisation de coût.

<u>Algorithm k-means</u>

Input : $T \in \mathbb{N}$, $\{x_i\}_{1\le i\le n}$

$\mu_j \longleftarrow$ random point from $\mathbb{R}^d$ for all $j \in [k]$

for $t = 1$ to T do :

① compute a Voroni partitioning $(C_1, \ldots, C_K)$ corresponding to cluster centre

② $\mu_j \longleftarrow \dfrac{1}{|\{i\in\mathbb{N} : x_i \in C_j\}|} \sum_{i=1}^{n} x_i \cdot \mathbb{1}_{\{x_i\in C_j\}}$

end for

Output : $(C_1, \ldots, C_K)$

2. (1 point) Soient $\mu_1, \dots, \mu_K \in \mathbb{R}^d$ fixés, $(C_1, \dots, C_K) \in \mathcal{P}$ le partitionnement de Voronoi associé et $(C'_1, \dots, C'_K) \in \mathcal{P}$ une partition quelconque. Montrer qu'alors

$$\sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \mathbf{1}_{x_i \in C_j} \leq \sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \mathbf{1}_{x_i \in C'_j}.$$

(2) D'après la définition du partitionnement de Voronoi

$$C_j = \left\{ x \in \mathbb{R}^d : \|x - \mu_j\|_{\ell_2}^2 \leq \|x - \mu_\ell\|_{\ell_2}^2, \forall \ell \in [K] \right\} / \bigcup_{i=1}^{j-1} C_i$$

alors

$\forall j \in [\![1,n]\!]$: $\forall x \in C_j$, on a

$$\|x - \mu_j\|_{\ell_2}^2 \leq \|x - \mu_\ell\|_{\ell_2}^2 \quad \forall \ell \in [\![1,n]\!]$$

donc

?

$$\|x - \mu_j\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x \in C_j\}} \leq \|x - \mu_\ell\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x \in C'_j\}}$$

alors $\displaystyle\sum_{j=1}^{K} \|x - \mu_j\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x \in C_j\}} \leq \sum_{\ell=1}^{K} \|x - \mu_\ell\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x \in C'_j\}}$

pour $(x_1, \dots, x_n)$ on a

$$\sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x_i \in C_j\}} \leq \sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - \mu_j\|_{\ell_2}^2 \cdot \mathbb{1}_{\{x \in C'_j\}}$$

3. (1 point) Soient $\{p_i\}_{1\le i\le n} \subset \mathbb{R}_+^*$ fixés. Expliciter :

$$\arg\min_{\mu\in\mathbb{R}^d} \sum_{i=1}^{n} p_i \|x_i - \mu\|_{\ell_2}^2 .$$

(3) Pour $\{p_i\}_{1\le i\le n}$ fixés, soit

$$\varphi(\mu) = \sum_{i=1}^{n} p_i \|x_i - \mu\|_{\ell_2}^2$$

on a $\varphi(\mu)$ est convexe

alors pour

$$\nabla\varphi(\hat{\mu}) = \sum_{i=1}^{n} 2\cdot p_i \cdot (\hat{\mu} - x_i) = 0$$

on a

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

4. (1 point) Quelles sont les différences entre le problème considéré ici et la méthode appelée *soft K-means* ?

(4) Pour le soft K-means

(1) $P_{ij}^{t} \sim \mathbb{P}(Y=j \mid X_i)$

(2) $\hat{\pi}^{t+1}$

(3) $\hat{\mu}_j^{t+1} \propto \sum_{i=1}^{n} P_{ij}^{t} \cdot X_i$

(4) $\hat{\Sigma}$

Pour le problème considéré

(1) $P_{ij}^{t} \sim \mathbb{1}_{\{x_i \in C_j\}}$

(2) $\mu_j^{t+1} \propto \sum_{i=1}^{n} P_{ij}^{t} X_i$

Exercice 2 (Modèle de classification, 6 points)

Rappels sur les lois de probabilité

- R est une variable aléatoire de Rademacher de paramètre $p \in]0,1[$ (on écrit $R \sim \mathcal{R}(p)$) si R est à valeurs dans $\{\pm 1\}$ telle que $\mathbb{P}(R=1)=p$.
- E est une variable aléatoire exponentielle (on écrit $E \sim \mathcal{E}$) si E est à valeurs dans $\mathbb{R}_+$ et de fonction de répartition $x \in \mathbb{R} \mapsto (1-e^{-x})\mathbf{1}_{\mathbb{R}_+}(x)$.

Soient (X,Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$, $f^\star : \mathbb{R}^d \to \mathbb{R}$ une fonction mesurable et ε une variable aléatoire réelle telle que $\varepsilon \,|\, X \sim \mathcal{E}$. On considère le modèle de classification :

$$Y = \operatorname{sign}\left(f^\star(X) + \varepsilon\right).$$

1. (1 point) Montrer que $Y|X \sim \mathcal{R}(p_X)$, avec p_X à déterminer.

1)
$$\mathbb{P}(Y=1|X) = \mathbb{P}(f^*(X)+\varepsilon > 0|X)$$
$$= \mathbb{P}(\varepsilon > -f^*(X)|X)$$
$$= 1-\mathbb{P}(\varepsilon \leq -f^*(X)|X)$$
$$= 1-(1-e^{f^*(X)})\cdot \mathbb{1}_{\mathbb{R}_+}(-f^*(X))$$

donc on a

$$p_X = 1-(1-e^{f^*(X)})\cdot \mathbb{1}_{\mathbb{R}_+}(-f^*(X))$$

et $\quad Y|X \sim \mathcal{R}(p_X)$

2. (1 point) Soit $x \in \mathbb{R}^d$. Exprimer $-\log\left(\mathbb{P}(Y=1 \,|\, X=x)\right)$ comme fonction de la variable $f^\star(x)$:

$$\ell_1(f^\star(x)) = -\log\left(\mathbb{P}(Y=1 \,|\, X=x)\right).$$

Représenter le graphe de ℓ_1. Est-ce une fonction convexe ?

(2)
$$\ell_1(f^*(x)) = -\log(\mathbb{P}(Y=1|X=x))$$
$$= -\log(1-(1-e^{f^*(x)})\cdot \mathbb{1}_{\mathbb{R}_+}(-f^*(x)))$$

Pour la fonction $\ell_1(x) = -\log(1-(1-e^{x})\cdot \mathbb{1}_{\mathbb{R}_+}(-x))$

$$= \begin{cases} -x, & x<0 \\ 0, & x \geq 0 \end{cases}$$

日期: /

on a la graphe

c'est une fonction convexe.

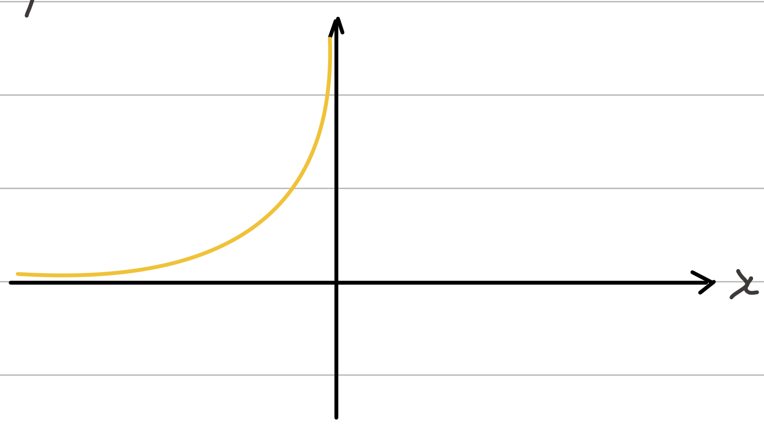
3. (1 point) Même question pour $\ell_{-1}(f^\star(x)) = -\log\left(\mathbb{P}(Y=-1 \,|\, X=x)\right)$.

(3) $\mathbb{P}(Y=-1|X=x) = (1-e^{f^*(x)})\cdot \mathbb{1}_{\mathbb{R}_+}(-f^*(x))$

$\ell_{-1}(f^*(x)) = -\log\left((1-e^{f^*(x)})\cdot \mathbb{1}_{\mathbb{R}_+}(-f^*(x))\right)$

pour la fonction $\ell_{-1}(x) = \begin{cases} \log\left(\frac{1}{1-e^x}\right) & \text{si } x < 0 \\ +\infty & \text{si } x \geqslant 0 \end{cases}$

alors la graphe



4. (2 points) On considère à présent le modèle linéaire

$$\forall x \in \mathbb{R}^d : \qquad f^\star(x) = \beta^{\star\top} x,$$

avec $\beta^\star \in \mathbb{R}^d$ inconnu. Comment obtenir un estimateur du maximum de vraisemblance de $\beta^\star$ à partir d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y) (expliciter le problème de minimisation associé à la log-vraisemblance avec ses éventuelles contraintes) ?

(4) Comme $X \sim \mathcal{E}$, alors $p(x) = e^{-x} \cdot \mathbb{1}_{\mathbb{R}^+}(x)$

on a

$$f_{(X,Y)}(x,y) = p(x) \cdot \mathbb{P}(Y=1 \mid X=x)^{\mathbb{1}_{\{y=1\}}} \cdot \mathbb{P}(Y=-1 \mid X=x)^{\mathbb{1}_{\{y=-1\}}}$$

donc

$$-\log f_{(X,Y)}(x,y) = -\log p(x) + \ell_1(f^*(x)) \cdot \mathbb{1}_{\{y=1\}} + \ell_{-1}(f^*(x)) \cdot \mathbb{1}_{\{y=-1\}}$$

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

$$-\log f_{(X,Y)}(X_i,Y_i) = -\log p(X_i) + \ell_1(\beta^{*T} X_i) \cdot \mathbb{1}_{\{Y_i=1\}} + \ell_{-1}(\beta^{*T} X_i) \cdot \mathbb{1}_{\{Y_i=-1\}}$$

alors le problème de minimisation est

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \sum_{i=1}^n -\log f_{(X,Y)}(X_i,Y_i)$$

$$\iff$$

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \sum_{i=1}^n \left(\ell_1(\beta^T X_i) \cdot \mathbb{1}_{\{Y_i=1\}} + \ell_{-1}(\beta^T X_i) \cdot \mathbb{1}_{\{Y_i=-1\}} \right)$$

$$\iff$$

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \sum_{i=1}^n -\log\left(1-(1-e^{\beta^T X_i}) \cdot \mathbb{1}_{\{\beta^T X_i \leq 0\}}\right) \cdot \mathbb{1}_{\{Y_i=1\}} - \log\left((1-e^{\beta^T X_i}) \cdot \mathbb{1}_{\{\beta^T X_i \leq 0\}}\right) \cdot \mathbb{1}_{\{Y_i=-1\}}$$

s.c. $\forall i \in [\![1,n]\!]$, $\mathbb{1}_{\{Y_i=-1\}} \cdot \beta^T X_i \leq 0$

5. (1 point) Déterminer le gradient de la fonction objectif introduite à la question précédente lorsqu'il existe.

(5) pour $\ell_1(\beta^T X_i) = \begin{cases} -\beta^T X_i & \text{si } \beta^T X_i \leq 0 \\ 0 & \text{si } \beta^T X_i > 0 \end{cases}$

$$\nabla_\beta \ell_1 = \begin{cases} -X_i & \text{si } \beta^T X_i \leq 0 \\ 0 & \text{si } \beta^T X_i > 0 \end{cases}$$

pour $\ell_{-1}(\beta^T X_i) = \begin{cases} -\log(1-e^{\beta^T X_i}) & \text{si } \beta^T X_i \leq 0 \\ +\infty & \text{si } \beta^T X_i > 0 \end{cases}$

$$\nabla_\beta \ell_{-1} = \begin{cases} \dfrac{X_i \cdot e^{\beta^T X_i}}{1-e^{\beta^T X_i}} & \text{si } \beta^T X_i \leq 0 \\ & \text{si } \beta^T X_i > 0 \end{cases}$$

Donc

$$\nabla_\beta \psi = \sum_{i=1}^{n} -X_i \cdot \mathbb{1}_{\{\beta^T X_i \leq 0\}} \cdot \mathbb{1}_{\{Y_i = 1\}} + \frac{X_i \cdot e^{\beta^T X_i}}{1-e^{\beta^T X_i}} \cdot \mathbb{1}_{\{\beta^T X_i \leq 0\}} \cdot \mathbb{1}_{\{Y_i = -1\}}$$

0

Exercice 3 (Approximation de noyau, 10 points)

Rappels généraux et de concentration

— Vecteur gaussien : un vecteur aléatoire gaussien X de dimension d, de moyenne $\mu \in \mathbb{R}^d$ et de matrice de covariance Σ symétrique définie positive (on notera $X \sim \mathcal{N}(\mu, \Sigma)$) admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, s'exprimant :

$$x \in \mathbb{R}^d \mapsto \frac{1}{\sqrt{2\pi}^d |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

— Trigonométrie :

$$\forall a, b \in \mathbb{R}, \quad \cos a \cos b = \frac{\cos(a+b) + \cos(a-b)}{2}.$$

— Inégalité de Markov : soit Z une variable aléatoire réelle positive presque sûrement. Alors

$$\forall t > 0, \quad \mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z^2]}{t^2}.$$

— Inégalité de Hoeffding : soit $(X_N)_{N\geq 1}$ une suite de variables aléatoires réelles indépendantes et bornées par $a > 0$, *i.e.* $|X_i| \leq a$ p.s. pour tout $i \in \mathbb{N}^*$. Alors en notant $S_N = \sum_{i=1}^N X_i$,

$$\forall t > 0, \quad \mathbb{P}(|S_N - \mathbb{E}[S_N]| \geq t) \leq 2e^{-\frac{t^2}{2Na^2}}.$$

Partie A

1. (1½ points) Soient $\gamma > 0$ et $\omega \sim \mathcal{N}(0, 2\gamma I_d)$. Montrer que pour tout $u \in \mathbb{R}^d$:

$$e^{\gamma\|u\|_{\ell_2}^2}\, \mathbb{E}[e^{i\omega^\top u}] = 1.$$

En déduire que

$$e^{-\gamma\|u\|_{\ell_2}^2} = \mathbb{E}[\cos(\omega^\top u)].$$

(1) $$\mathbb{E}[e^{i w^\top u}] = \int_{\mathbb{R}^d} e^{i w^\top u} \cdot \frac{1}{\sqrt{2\pi}^d \sqrt{2\gamma}} \cdot e^{-\frac{1}{2} w^\top \cdot \frac{1}{2\gamma} \cdot w}\, dw$$

$$= \frac{1}{\sqrt{2\pi}^d\sqrt{2\gamma}} \cdot \int_{\mathbb{R}^d} e^{-\frac{1}{4\gamma}\cdot(w^\top w - 4\gamma \cdot i \cdot w^\top u)}\, dw$$

$$= \frac{1}{\sqrt{2\pi}^d\sqrt{2\gamma}} \cdot e^{-\gamma u^\top u} \cdot \int_{\mathbb{R}^d} e^{-\frac{\|w - 2\gamma \cdot i \cdot u\|_{\ell_2}^2}{4\gamma}}\, dw$$

$$= e^{-\gamma \cdot \|u\|_{\ell_2}^2} \cdot \int_{\mathbb{R}^d} \frac{1}{\sqrt{2\pi}^d \sqrt{2\gamma}} \cdot e^{-\frac{1}{2}\cdot(w - 2\gamma i u)^\top \cdot (2\gamma \cdot I_d)^{-1} \cdot (w - 2\gamma \cdot i \cdot u)}\, dw$$

$$= e^{-\gamma \cdot \|u\|_{\ell_2}^2}$$

donc

$$e^{\gamma \cdot \|u\|_{\ell_2}^2} \cdot \mathbb{E}[e^{i w^\top u}] = 1$$

日期:　　　/

D'après la formule d'Euler,

$$\mathbb{E}[e^{iw^T u}] = \mathbb{E}[\cos(w^T u) + i\cdot\sin(w^T u)]$$
$$= \mathbb{E}[\cos(w^T u)] + i\cdot\mathbb{E}[\sin(w^T u)]$$

et comme $\mathbb{E}[e^{iw^T u}] = e^{-\gamma\cdot\|u\|_{\ell_2}^2} \in \mathbb{R}$

donc $\mathbb{E}[\sin(w^T u)] = 0$

on a $\mathbb{E}[e^{iw^T u}] = \mathbb{E}[\cos(w^T u)] = e^{-\gamma\cdot\|u\|_{\ell_2}^2}$

2. (1 point) Soit $\varphi \sim \mathcal{U}([0,\pi])$ indépendante de ω. Montrer, par intégration, que pour tout $x, y \in \mathbb{R}^d$:

$$\mathbb{E}[\cos(\omega^\top(x-y))] = 2\,\mathbb{E}[\cos(\omega^\top x+\varphi)\cos(\omega^\top y+\varphi)].$$

(2) $2\mathbb{E}[\cos(w^T x+\varphi)\cdot\cos(w^T y+\varphi)]$

$$= \mathbb{E}[\cos(w^T(x-y)) + \cos(w^T(x+y)+2\varphi)]$$
$$= \mathbb{E}[\cos(w^T(x-y))] + \mathbb{E}[\cos(w^T(x+y)+2\varphi)]$$

Donc il faut montrer

$$\forall x, y \in \mathbb{R}^d : \mathbb{E}[\cos(w^T(x+y)+2\varphi)] = 0$$

on a

$$\mathbb{E}[\cos(w^T(x+y)+2\varphi)] = \int_{\mathbb{R}^d}\int_0^{\pi} \cos(w^T(x+y)+\varphi)\cdot\frac{1}{\pi}\,d\varphi\cdot f_N(w)\,dw$$
$$= \int_{\mathbb{R}^d} \sin(w^T(x+y)+\varphi)\Big|_0^{\pi}\cdot\frac{1}{\pi}\cdot f_N(w)\,dw$$
$$= -\frac{2}{\pi}\cdot\mathbb{E}[\sin(w^T(x+y))]$$
$$= 0$$

3. (1½ points) Soient $(\omega_N)_{N\geq 1}$ une suite de vecteurs aléatoires i.i.d. de même loi que ω et indépendants de $(\varphi_N)_{N\geq 1}$ i.i.d. de même loi que φ :

$$(\omega_N)_{N\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, 2\gamma I_d) \perp\!\!\!\perp (\varphi_N)_{N\geq 1} \overset{i.i.d.}{\sim} \mathcal{U}([0,\pi]).$$

Pour un entier $N > 1$ fixé, et $x, y \in \mathbb{R}^d$ quelconques, on s'intéresse à l'approximation :

$$e^{-\gamma\|x-y\|_{\ell_2}^2} \approx \frac{2}{N}\sum_{i=1}^{N} \cos(\omega_i^\top x + \varphi_i)\cos(\omega_i^\top y + \varphi_i).$$

Quel est l'intérêt de celle-ci pour l'apprentissage automatique (penser aux aspects algorithmiques lors des phases d'apprentisage et de prédiction) ?

(3) Avec cette approximation, on peut calculer l'approximation de noyau gaussien dans un plus facile manière.

Pour les algorithms avec kernel trick, on peut calculer numériquement $h^* = \sum_{i=1}^{n} \alpha_i^* \cdot k(\cdot, X_i)$ plus efficient.

① Amélioration de l'efficacité de calcul

② Réduction de l'usage de mémoire

③ Simplification du modèle tout en conservant la précision

④ Adaptation aux tâches d'apprentissage en ligne ou en temps réel

4. (1 point) Montrer que pour tout $\epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{2}{N}\sum_{i=1}^{N}\cos(\omega_i^\top x+\varphi_i)\cos(\omega_i^\top y+\varphi_i)-e^{-\gamma\|x-y\|_{\ell_2}^2}\right|\geq\epsilon\right)\leq 2e^{-\frac{N\epsilon^2}{8}}.$$

(4) D'après 3, soit $X_i = \frac{2}{N}\cdot\cos(\omega_i^T x+\varphi_i)\cdot\cos(\omega_i^T y+\varphi_i)$ $\forall i$

et soit $S_N = \sum_{i=1}^{N} X_i$, on a $E[S_N] = e^{-\gamma\|x-y\|_{\ell_2}^2}$

et on a $|X_i| \leq \frac{2}{N}$.

D'après l'inégalité de Hoeffding, on a

$$\mathbb{P}(|S_N - E[S_N]| \geq \varepsilon) \leq 2\cdot e^{-\frac{\varepsilon^2}{2\cdot N\cdot\frac{4}{N^2}}} = 2\cdot e^{-\frac{N\cdot\varepsilon^2}{8}}$$

5. (2 points) Soient $\mathcal{S}\subset\mathbb{R}^d$ un ensemble de cardinal $n > 1$ fixé et $\delta\in]0,1[$. Montrer qu'avec probabilité au moins $1-\delta$:

$$\forall x,y\in\mathcal{S},\quad\left|\frac{2}{N}\sum_{i=1}^{N}\cos(\omega_i^\top x+\varphi_i)\cos(\omega_i^\top y+\varphi_i)-e^{-\gamma\|x-y\|_{\ell_2}^2}\right|\leq\sqrt{\frac{16}{N}\log\left(\frac{n+1}{\sqrt{\delta}}\right)}.$$

(5) D'après (4), $\forall x,y\in S$, soit

$$S_N = \sum_{i=1}^{N}\frac{2}{N}\cdot\cos(\omega_i^T x+\varphi_i)\cdot\cos(\omega_i^T y+\varphi_i)$$

$$E[S_N] = e^{-\gamma\|x-y\|_{\ell_2}^2}$$

alors

$$\mathbb{P}\left(|S_N - E[S_N]| \leq \sqrt{\frac{16}{N}\log\left(\frac{n+1}{\sqrt{\delta}}\right)}\right)$$

$$= 1-\mathbb{P}\left(|S_N - E[S_N]| > \sqrt{\frac{16}{N}\log\left(\frac{n+1}{\sqrt{\delta}}\right)}\right)$$

$$\geq 1-2\cdot\exp\left(-\frac{N}{8}\cdot\frac{16}{N}\cdot\log\left(\frac{n+1}{\sqrt{\delta}}\right)\right)$$

$$= 1-2\cdot\exp\left(\log\left(\frac{\delta}{(n+1)^2}\right)\right)$$

$$= 1-\frac{2}{(n+1)^2}\cdot\delta$$

$$\geq 1-\delta$$

Partie B

On souhaite à présent établir une borne similaire à celle de la question précédente mais pour un ensemble non-dénombrable.

Soient $\mathcal{X} \subset \mathbb{R}^d$ un espace borné par $\frac{1}{8}$: $\sup_{x \in \mathcal{X}} \|x\|_{\ell_2} \leq \frac{1}{8}$; et $r > 0$ un rayon quelconque.

On admet qu'il existe $T \leq r^{-d}$ boules de rayons r recouvrant $\mathcal{U} = \{x - y : x, y \in \mathcal{X}\}$ et on note $U_1, \ldots, U_T$ leurs centres.

On appelle f_N l'erreur d'approximation, définie pour tout $u \in \mathcal{U}$ par :

$$f_N(u) = \frac{2}{N} \sum_{i=1}^{N} \cos(\omega_i^\top x_u + \varphi_i) \cos(\omega_i^\top y_u + \varphi_i) - e^{-\gamma \|u\|_{\ell_2}^2},$$

où $x_u, y_u \in \mathcal{X}$ sont tels que $u = x_u - y_u$, et on admet que f_N est L_N-lipschitzienne ($L_N > 0$, p.s.) avec $\sigma^2 = \mathbb{E}[L_N^2] \leq \mathbb{E}[\|\omega\|_{\ell_2}^2]$.

1. (1 point) Montrer que $\sigma^2 \leq \mathrm{tr}(\mathbb{V}(\omega)) = 2d\gamma$.

(1) $\sigma^2 = E[L_N^2] \leq E[\|\omega\|_{\ell_2}^2] = E[\sum_{i=1}^{d} \omega_i^2]$

$= \sum_{i=1}^{d} E[\omega_i^2 - E[\omega_i]^2]$

$= \sum_{i=1}^{d} \mathrm{Var}[\omega_i]$

$= \mathrm{tr}(\mathbb{V}(\omega))$

$= 2d\gamma$

2. (1 point) Soient $u \in \mathcal{U}$, $i \in [\![1, T]\!]$ tel que $\|u - U_i\|_{\ell_2} \leq r$ et $\epsilon > 0$. Montrer que :

$$\left[|f_N(U_i)| \leq \frac{\epsilon}{2} \text{ et } L_N \leq \frac{\epsilon}{2r}\right] \implies [|f_N(u)| \leq \epsilon].$$

(2) $\forall u \in \mathcal{U}$, il existe une boule de rayon r, de centrale U_i tel que $u \in B(U_i, r)$, donc

$|f_N(u)| = |f_N(u) - f_N(U_i) + f_N(U_i)|$

$\leq |f_N(u) - f_N(U_i)| + |f_N(U_i)|$

$\leq L_N \cdot |u - U_i| + \frac{\varepsilon}{2}$

$\leq \frac{\varepsilon}{2r} \cdot r + \frac{\varepsilon}{2}$

$\leq \varepsilon$

3. (1 point) En déduire que :

$$\mathbb{P}\left(\sup_{u\in\mathcal{U}}|f_N(u)|>\epsilon\right)\le\sum_{i=1}^{T}\mathbb{P}\left(|f_N(U_i)|>\frac{\epsilon}{2}\right)+\mathbb{P}\left(L_N>\frac{\epsilon}{2r}\right).$$

(3) $\mathbb{P}(\sup_{u\in\mathcal{U}}|f_N(u)|>\varepsilon) = 1-\mathbb{P}(\sup_{u\in\mathcal{U}}|f_N(u)|\leq\varepsilon)$

et

$\sup_{u\in\mathcal{U}}|f_N(u)|\leq\varepsilon \iff \forall i\in[\![1,T]\!],\ \forall u\in\mathcal{U}_i,\ |f_N(u)|\leq\varepsilon$

et d'après (2)

$\forall i\in[\![1,T]\!],\ |f_N(U_i)|\leq\frac{\varepsilon}{2},\ L_N\leq\frac{\varepsilon}{2r} \implies \forall i\in[\![1,T]\!],\ \forall u\in\mathcal{U}_i,\ |f_N(u)|\leq\varepsilon$

donc

$\forall i\in[\![1,T]\!],\ |f_N(U_i)|\leq\frac{\varepsilon}{2},\ L_N\leq\frac{\varepsilon}{2r} \implies \sup_{u\in\mathcal{U}}|f_N(u)|\leq\varepsilon$

alors

$\sup_{u\in\mathcal{U}}|f_N(u)|>\varepsilon \implies (\forall i\in[\![1,T]\!],|f_N(U_i)|>\frac{\varepsilon}{2})\cup(L_N>\frac{\varepsilon}{2r})$

on a

$\mathbb{P}(\sup_{u\in\mathcal{U}}|f_N(u)|>\varepsilon)\leq\sum_{i=1}^{T}\mathbb{P}(|f_N(U_i)|>\frac{\varepsilon}{2})+\mathbb{P}(L_N>\frac{\varepsilon}{2r})$

4. (2 points (bonus))

$$\mathbb{P}\left(\sup_{u\in\mathcal{U}}|f_N(u)|>\epsilon\right)\le\frac{4r^2\sigma^2}{\epsilon^2}+\frac{2}{r^d}e^{-\frac{N\epsilon^2}{32}}.$$

(4) D'après partie A (4)

$\mathbb{P}(|f_N(U_i)|>\frac{\varepsilon}{2})\leq 2\cdot e^{-\frac{N}{8}\cdot\frac{\varepsilon^2}{4}} = 2\cdot e^{-\frac{N\varepsilon^2}{32}}$

D'après l'inégalité de Markov

$\mathbb{P}(L_N>\frac{\varepsilon}{2r})\leq\frac{E[L_N^2]}{\frac{\varepsilon^2}{4r^2}} = \frac{4r^2\sigma^2}{\varepsilon^2}$

Donc

$$\mathbb{P}\left(\sup_{u\in\mathcal{U}} |f_N(u)| > \varepsilon\right) \leq \sum_{i=1}^{T} 2\cdot e^{-\frac{N\varepsilon^2}{32}} + \frac{4r^2\sigma^2}{\varepsilon^2}$$

$$= T\cdot 2\cdot e^{-\frac{N\varepsilon^2}{32}} + \frac{4r^2\sigma^2}{\varepsilon^2}$$

$$\leq \frac{2}{r^d}\cdot e^{-\frac{N\varepsilon^2}{32}} + \frac{4r^2\sigma^2}{\varepsilon^2}$$

5. (2 points (bonus)) En remarquant que pour tout $a, b > 0$,

$$ar^2 + \frac{b}{r^d} = 2a^{\frac{d}{d+2}} b^{\frac{2}{d+2}},$$

lorsque $r = \left(\frac{b}{a}\right)^{\frac{1}{d+2}}$, montrer que pour $\epsilon \in]0, \sigma]$:

$$\mathbb{P}\left(\sup_{x,y\in\mathcal{X}} \left| \frac{2}{N}\sum_{i=1}^{N} \cos(\omega_i^\top x + \varphi_i)\cos(\omega_i^\top y + \varphi_i) - e^{-\gamma\|x-y\|_{\ell_2}^2} \right| > \epsilon\right) \leq 2^4 d\frac{\gamma}{\epsilon^2} e^{-\frac{N\epsilon^2}{16(d+2)}}.$$

(5) D'après (4)

$$\mathbb{P}\left(\sup_{x,y\in\mathcal{X}} \left|\frac{2}{N}\sum_{i=1}^{N} \cos(w_i^T x + \varphi_i)\cdot\cos(w_i^T y + \varphi_i) - e^{-\gamma\|x-y\|_{\ell_2}^2}\right| > \varepsilon\right)$$

$$= \mathbb{P}\left(\sup_{u\in\mathcal{U}} |f_N(u)| > \varepsilon\right)$$

$$\leq \frac{4\sigma^2}{\varepsilon^2}\cdot r^2 + 2\cdot e^{-\frac{N\varepsilon^2}{32}}\cdot\frac{1}{r^d} \quad ①$$

lorsque $r = \left(2\cdot e^{-\frac{N\varepsilon^2}{32}}\cdot\frac{\varepsilon^2}{4\sigma^2}\right)^{\frac{1}{d+2}} = \left(\frac{\varepsilon^2}{2\sigma^2}\cdot e^{-\frac{N\varepsilon^2}{32}}\right)^{\frac{1}{d+2}}$

on a

$$① = 2\cdot\left(\frac{4\sigma^2}{\varepsilon^2}\right)^{\frac{d}{d+2}}\cdot\left(2\cdot e^{-\frac{N\varepsilon^2}{32}}\right)^{\frac{2}{d+2}}$$

$$\leq 2\cdot\frac{4\sigma^2}{\varepsilon^2}\cdot 2^{\frac{2}{d+2}}\cdot e^{-\frac{N\varepsilon^2}{16(d+2)}}$$

$$\leq 2^4\cdot\frac{d\gamma}{\varepsilon^2}\cdot 1\cdot e^{-\frac{N\varepsilon^2}{16(d+2)}}$$

日期: /