Aix*Marseille université

MASTER 2 INFORMATIQUE

PARCOURS IAAA, UE MIA

Année 2021-22

DM 2 : Statistique et Optimisation

Exercice 1 Analyse des propriétés basiques d'un estimateur. (Faisable après la cinquième séance)

Supposons qu'on observe un échantillon $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$, i.i.d. de distribution P. Soit $d : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$ une fonction mesurant la 'dissimilarité' entre deux points de $\mathbf{R}^d$. On suppose que d est symmétrique, c'est à dire que $d(x_1, x_2) = d(x_2, x_1)$ pour tout choix de x_1, x_2. Nous cherchons à estimer la dissimilarité moyenne entre deux points tirés aléatoirement et indépendamment suivant la distribution P :

$$\delta(P, d) := \mathbf{E}_{a,b \sim P \otimes P}[d(a, b)]$$

sur la base de $x_1, x_2, ..., x_n$ (la notation $a, b \sim P \otimes P$ signifie que a et b sont deux échantillons tirés indépendamment de la loi P). On suppose que $\mathbf{E}_{a,b \sim P \otimes P}[d(a, b)^2] < +\infty$.

Considérons l'estimateur :

$$\hat{\delta}(x_1, x_2, \ldots, x_n) := \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(x_i, x_j).$$

1. Quel est le biais de $\hat{\delta}$?
2. Quelle est la variance de $\hat{\delta}$? On l'exprimera en fonction de $\sigma_1 = \mathrm{Var}_{x_1 \sim P} \mathbf{E}_{x_2 \sim P}[d(x_1, x_2)]$ et $\sigma_2 = \mathrm{Var}_{(x_1, x_2) \sim P \otimes P}[d(x_1, x_2)]$.
3. Prouver l'inégalité de Markov : soit X est un variable aléatoire réelle positive, de moyenne finie μ et soit t un réel strictement positif, alors :
$$p(X \geq t) \leq \frac{\mu}{t}.$$
4. Utiliser l'inégalité de Markov pour prouver l'inégalité de Chebyshev : soit X est un variable aléatoire réelle positive, de moyenne finie μ et de variance finie σ^2 et soit t un réel strictement positif, alors :
$$p(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$
5. Utiliser l'inégalité de Chebyshev et les résultats des deux premières questions pour montrer que $\hat{\delta}$ est un estimateur faiblement consistant de δ.

Exercice 2 Intervalles de confiance. (Faisable après la cinquième séance)

On considère des variables aléatoires rélles X_1, X_2,..., X_n i.i.d. de loi P et de variance :

$$\mathrm{Var}(P) = \sigma^2 < +\infty.$$

On suppose de plus que X_1, X_2,..., X_n prennent leur valeurs dans l'intervalle $[0, 1]$.

1. En vous appuyant sur le théorème central limite, donnez l'expression d'un intervalle de confiance asymptotique de niveau $1 - \alpha$ pour la moyenne empirique $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$ et prouvez que l'intervalle proposé est effectivement asymptotiquement de niveau $1 - \alpha$.
2. En vous appuyant sur l'inégalité de concentration de Hoeffding, donnez l'expression d'un intervalle de confiance (non-asymptotique) de niveau $1-\alpha$ pour $\hat{\mu}$ et prouvez que l'intervalle proposé est de niveau $1 - \alpha$.
3. Comparez la largeur des intervalles obtenus par les deux méthodes et commentez.

Exercice 3 Calculs de gradients. (Faisable après la sixième séance)

Calculez le gradient des fonctions suivantes (en détaillant vos calculs).

1. $$f_1 : \left| \begin{array}{l} \mathbf{R}^2 \to \mathbf{R} \\ x, y \mapsto 2\cos(x^2 + y) - 4x \end{array} \right.$$

2. $$f_2 : \left| \begin{array}{l} \mathbf{R}^2 \to \mathbf{R} \\ x, y \mapsto \exp(-x^2 - y^2) \end{array} \right.$$

3. $$f_{3,A} : \left| \begin{array}{l} \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R} \\ x, y \mapsto x^T A y \end{array} \right. ,$$

   for an m by n matrix with real coefficients A.

4. $$f_{4,x,y} : \left| \begin{array}{l} \mathbf{R}^{m\times n} \to \mathbf{R} \\ A \mapsto x^T A y \end{array} \right. ,$$

   for x and y two column matrices with real coefficients of size respectively m and n.

5. $$f_{5,A,Y} : \left| \begin{array}{l} \mathbf{R}^{m\times n} \to \mathbf{R} \\ X \mapsto \|Y - AX\|_2 \end{array} \right. ,$$

for Y a l by n matrix with real coefficients and A an l by m matrix with real coefficients, and where the matrix norm used is the spectral norm.

Exercice 4 Estimation par maximum de vraisemblanc des paramètres d'une loi Gaussienne multivariée. (Faisable après la sixième séance)

Supposons qu'on observe un échantillon i.i.d. $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$ de loi gaussienne multivariée $\mathcal{N}(\mu^*, \Sigma^*)$, pour un vecteur $\mu \in \mathbf{R}^d$ et une matrice $\Sigma^* \in \mathbf{S}_d$, où $\mathbf{S}_d$ est l'ensemble des matrices à coefficients réels symétriques définies positives de taille d pas d. On cherche à estimer μ^* et Σ^* à partir de l'observation de $x_1, x_2, \ldots, x_n$.

On définit la *vraisemblance* d'un couple de paramètres $(\mu, \Sigma) \in \mathbf{R}^d \times \mathbf{S}_d$ comme :

$$\ell(\mu, \Sigma) := p(x_1, x_2, \ldots, x_n; \mu, \Sigma),$$

où p corresponds à la densité de probabilité de $x_1, x_2, \ldots, x_n$.

On considère l'estimateur du *maximum de vraisemblance* pour μ^*, Σ^* défini par

$$\hat{\mu}, \hat{\Sigma} \in \arg \max_{(\mu,\Sigma) \in \mathbf{R}^d \times \mathbf{S}_d} \ell(\mu, \Sigma).$$

1. Montrer que
$$\hat{\mu}, \hat{\Sigma} \in \arg \max_{(\mu,\Sigma) \in \mathbf{R}^d \times \mathbf{S}_d} \log(\ell(\mu, \Sigma)).$$
2. Donner une expression (la plus simple que vous pouvez) pour $\log(\ell(\mu, \Sigma)))$ en utilisant la formule donnant la densité d'une loi gaussienne multivariée non dégénérée.
3. On admet que $\log(\ell(\mu, \Sigma)))$ est de classe C^1 et admet un unique maximum sur $\mathbf{R}^d \times \mathbf{S}_d$ en un point où son gradient s'annule. Calculer une expression explicite pour $(\hat{\mu}, \hat{\Sigma})$ en fonction de $x_1, x_2, \ldots, x_n$.
4. Montrer que le couple $(\hat{\mu}, \hat{\Sigma})$ ainsi obtenu forme une statistique suffisante pour le couple de paramètres (μ^*, Σ^*).

Exercice 5 Analyse de l'algorithme de descente de gradient pour une fonction de coût lisse et fortement convexe. (Faisable après la sixième séance)

Définitions :

— **Fonction L-lisse.** Soit L un nombre réel positif, une fonction $f : \mathbf{R}^n \to \mathbf{R}$ est dite L-lisse si et seulement si elle est de classe C^1 et que son gradient est L-lipschitzien, c'est à dire que pour tous vecteurs x, y de $\mathbf{R}^n$: $\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$.

— **Fonction μ-fortement convexe.** Soit μ un nombre réel strictement positif, une fonction différentiable $f : \mathbf{R}^n \to \mathbf{R}$ est dite μ-fortement convexe si et seulement si pour tous vecteurs x, y de $\mathbf{R}^n$, $f(x) \geq f(y) + \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|_2^2$.

Soit $r : R^n \to \mathbf{R}$ une fonction à minimiser. On suppose que r est L-lisse et μ-fortement convexe pour $L \geq 0$ et $\mu > 0$.

On note x^* l'unique minimum de la fonction r et on considère les valeurs successives $x_1, x_2, \ldots, x_t, \ldots$ obtenues par descente du gradient de r avec un pas fixe $\gamma = \frac{1}{L}$ en partant de $x_0 \in \mathbf{R}^n$.

1. Montrez que $1 - x \leq \exp(-x)$ pour tout réel x
2. Soit $x \in \mathbf{R}^n$. Montrez que la fonction :

$$g_x : \left| \begin{array}{l} \mathbf{R}^n \times \mathbf{R} \to \mathbf{R} \\ z \mapsto r(x) + \nabla r(x)^T(z-x) + \frac{\mu}{2}\|z-x\|_2^2 \end{array} \right.$$

   est strictement convexe et donnez une expression explicite pour le point z^* où elle atteint son minimum.
3. En utilisant le résultat de la question précédente, montrez que pour $x \in \mathbf{R}^n$, $\|\nabla r(x)\|_2^2 \geq 2\mu\left(r(x) - r(x^*)\right)$.
4. Donnez une interprétation verbale de la propriété que vous avez établi à la question précédente.
5. En utilisant les résultats des questions 1 et 3 montrez que pour tout entier naturel t :

$$r(x_t) - r(x^*) \leq \exp\left(-t\mu/L\right)\left(r(x_0) - r(x^*)\right).$$

6. Que nous apprend ce résultat sur la méthode de descente de gradient à pas fixe pour une fonction de coût lisse et fortement convexe ?

Exercice 6 (Bonus) Weyl's inequality. (Faisable après la sixième séance)

Nous allons prouver un résultat qui peut être utile, par exemple, pour analyser les propriétés statistiques de l'estimateur usuel utilisé dans le cadre d'une analyse en composantes principales.

Considérons une matrice symmétrique M à coefficients réels et de taille d par d à laquelle on ajoute une perturbation, représentée par une autre matrice symétrique P à coefficients réels et de taille d par d. On cherche à trouver des conditions sous lesquelles on peut garantir que les valeurs propres de M et celles de $M' = M + P$ sont proches.

Pour toute matrice symmétrique A à coefficients réels et de taille d par d, on note $\lambda_1(A), \lambda_2, \ldots, \lambda_d(A)$ les valeurs propres de A triées par ordre décroissant.

1. Montrer que :

$$|\lambda_1(M') - \lambda_1(M)| \leq \|P\|.$$

   (Indice : nous avons défini et prouvé certaines propriétés de la norme matricielle $\|.\|_2$ en cours.)

2. Soit $i \in \{2, 3, \ldots, d\}$. Notons $\mathcal{E}_i^d$ l'ensemble de tous les sous-espaces de dimension i de $\mathbf{R}^d$. Montrer que pour toute matrice symmétrique A à coefficients réels et de taille d par d, la i-ième valeur propre de A est donnée par :

$$\lambda_i(A) = \min_{E \in \mathcal{E}_i^d} \max_{v \in E^\perp \cap \mathcal{S}^{d-1}} v^T A v,$$

   où $E^\perp$ est l'ensemble des vecteurs de $\mathbf{R}^d$ orthogonaux à tous les éléments de E et $\mathcal{S}^{d-1}$ est l'ensemble des vecteurs de $\mathbf{R}^d$ de norme euclidienne 1.

3. Montrer que :

$$\max_{i \in \{1,2,\ldots,d\}} |\lambda_i(M') - \lambda_i(M)| \leq \|P\|_2.$$