PROGRAMME INSTITUTS ET INITIATIVES

Appel à projet - campagne 2021
Proposition de projet de recherche doctoral (PRD)
SCAI - Sorbonne Center of Artificial Intelligence

Intitulé du projet de recherche doctoral (PRD) : Concentration de l'estimateur des plus proches voisins.

Directeur de thèse porteur du projet (titulaire d'une HDR) :

NOM : Guyader
Prénom : Arnaud
Titre : Professeur des Universités
E-mail : arnaud.guyader@sorbonne-universite.fr
Adresse professionnelle :
Sorbonne Université
Bureau 211, Tour 15-25, Boîte 158
4, Place Jussieu, 75005 Paris
Phone : +(33) 1 44 27 26 29
Unité de recherche :
Laboratoire de Probabiliteés, Statistique et Modeélisation (LPSM), UMR 8001
École Doctorale de rattachement :
ED386-Sciences Mathématiques Paris Centre
Doctorant.e.s actuellement encadré.e.s : 1 (50%, soutenance en 2021).

Co-encadrant.e :

NOM : Ben-Hamou
Prénom : Anna
Titre : Maître de conférences
E-mail : anna.ben-hamou@upmc.fr
Unité de recherche :
Laboratoire de Probabilités, Statistique et Modélisation (LPSM), UMR 8001
École Doctorale de rattachement :
ED386-Sciences Mathématiques Paris Centre
Doctorant.e.s actuellement encadré.e.s : 0.

Cotutelle internationale : Non.

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ? Oui, ISCD - Institut des Sciences du calcul & des Données.

Description du projet de recherche doctoral :

Soit $(\mathbf{X}, Y)$ un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, avec $\mathbf{X} \sim \mu$ où μ est supposée à densité, et où $\mathbf{E}[Y^2] < \infty$. On cherche à comprendre comment la variable réponse Y dépend du vecteur $\mathbf{X}$, i.e. à trouver une fonction mesurable $g : \mathbb{R}^d \to \mathbb{R}$ telle que $g(\mathbf{X})$ soit une bonne approximation de Y. Le critère généralement retenu pour mesurer la qualité de l'estimation est celui du risque quadratique $\mathbf{E}[(Y - g(\mathbf{X}))^2]$. Un optimum est alors donné par la fonction de régression r donnée par

$$\forall x \in \mathbb{R}^d\,,\ r(x) = \mathbf{E}[Y \mid \mathbf{X} = x]\,,$$

qui vérifie

$$\mathbf{E}[(Y - r(\mathbf{X}))^2] = \inf_g \mathbf{E}[(Y - g(\mathbf{X}))^2]\,,$$

où l'infimum est pris sur toutes les fonctions mesurables $g : \mathbb{R}^d \to \mathbb{R}$ telles que $\mathbf{E}[g(\mathbf{X})^2] < \infty$. En pratique, la loi du couple $(\mathbf{X}, Y)$ est inconnue et l'on ne peut donc pas espérer pouvoir calculer $r(\mathbf{X})$. Cependant, si l'on dispose d'observations i.i.d. $\mathcal{D}_n = ((\mathbf{X}_i, Y_i))_{1 \le i \le n}$ de même loi que $(\mathbf{X}, Y)$, alors on peut chercher à estimer la fonction de régression à partir de l'échantillon $\mathcal{D}_n$. Notons qu'il s'agit d'un problème non-paramétrique : on ne suppose pas que la fonction r peut être décrite avec un nombre fini de paramètres, comme c'est le cas en régression linéaire.

Soit $x \in \mathbb{R}^d$. Le principe de l'estimateur des plus proches voisins est d'approcher $r(x)$ par une moyenne pondérée des valeurs Y_i, en mettant une pondération d'autant plus forte que $\mathbf{X}_i$ est proche de x. Plus précisément, notons

$$\Big(\mathbf{X}_{(1)}(x), Y_{(1)}(x)\Big) \ldots, \Big(\mathbf{X}_{(n)}(x), Y_{(n)}(x)\Big)$$

le réarrangement des observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ par ordre croissant des distances $\|\mathbf{X}_i - x\|$:

$$\|\mathbf{X}_{(1)}(x) - x\| \le \cdots \le \|\mathbf{X}_{(n)}(x) - x\|\,.$$

(Comme la loi de $\mathbf{X}$ est à densité, toutes ces inégalités sont strictes presque sûrement). L'estimateur des plus proches voisins (PPV) est alors défini par

$$r_n(x) = \sum_{i=1}^{n} w_i Y_{(i)}(x)\,, \tag{1}$$

où $(w_1, \ldots, w_n) \in [0,1]^n$ est une suite (déterministe) de poids telle que $\sum_{i=1}^n w_i = 1$. Généralement, on suppose que cette suite est décroissante : $w_1 \ge \cdots \ge w_n$. Un exemple important est celui de poids uniforme sur $[\![1, k]\!]$ pour $k \in [\![1, n]\!]$: $w_i = \frac{1}{k}\mathbb{1}_{\{i \le k\}}$. On

parle alors d'estimateur des k plus proches voisins (k-PPV).

Il existe de nombreux résultats asymptotiques pour cet estimateur. Par exemple, on dit que r_n est universellement L^p-consistant (pour $p \geq 1$) si pour toute loi de $(\mathbf{X}, Y)$ avec $\mathbf{E}[|Y|^p] < \infty$, on a

$$\mathbf{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|^p\right] \underset{n\to\infty}{\longrightarrow} 0\,,$$

où l'espérance est prise à la fois sur $\mathcal{D}_n$ et sur $\mathbf{X}$ (indépendant de $\mathcal{D}_n$). Le théorème de Stone (1977) implique que l'estimateur r_n défini en (1) avec (w_i) décroissante est universellement L^p-consistant si et seulement si il existe une suite d'entiers $(k_n)_{n\geq 1}$ telle que

1. $k_n \to +\infty$ et $\frac{k_n}{n} \to 0$;
2. $\sum_{i>k_n} w_i \to 0$;
3. $w_1 \to 0$.

En particulier l'estimateur des k_n plus proches voisins est universellement L^p-consistant ssi $k_n \to +\infty$ et $\frac{k_n}{n} \to 0$. On peut aussi montrer (Devroye, 1982) que sous certaines hypothèses, l'estimateur r_n est fortement ponctuellement consistant :

$$r_n(x) \xrightarrow{\text{p.s.}} r(x)\,, \quad \text{pour } \mu\text{-presque tout } x \in \mathbb{R}^d.$$

Sous des hypothèses plus fortes sur la loi de $(\mathbf{X}, Y)$ (en supposant en particulier que le support S de μ est compact) et sur la suite (w_i), on peut même montrer (Devroye, 1978) la consistance forte uniforme :

$$\sup_{x\in S} |r_n(x) - r(x)| \xrightarrow{\text{p.s.}} 0\,.$$

Pour une très bonne présentation de l'estimateur des PPV et de ses propriétés asymptotiques, voir Biau and Devroye [3], Györfi et al. [10], et Devroye et al. [8].

Il existe cependant assez peu de résultats sur les propriétés non-asymptotiques de cet estimateur. Le but de cette thèse est d'obtenir de bonnes inégalités de concentration exponentielles pour $r_n(x)$, autour de son espérance $\mathbf{E}[r_n(x)]$. Si l'on sait d'autre part contrôler le biais $|\mathbf{E}[r_n(x)] - r(x)|$, alors on obtiendra une borne non-asymptotique sur $|r_n(x) - r(x)|$, valable avec grande probabilité. Pour un contrôle uniforme, nous chercherons aussi à obtenir une borne sur $\mathbf{P}\left(\sup_{x\in S} |r_n(x) - r(x)| \geq \varepsilon\right)$. Enfin, nous proposons aussi d'étudier une variante de l'estimateur des k plus proches voisins, l'estimateur des k plus proches voisins mutuels, introduit par Guyader and Hengartner [9] et donné par

$$m_n(x) = \frac{1}{|\mathcal{M}_n(x)|} \sum_{i,\, \mathbf{X}_i \in \mathcal{M}_n(x)} Y_i\,,$$

où $\mathcal{M}_n(x)$ est l'ensemble des plus proches voisins mutuels de x, i.e. l'ensemble des points $\mathbf{X}_i$ qui non seulement appartiennent aux k plus proches voisins de x mais sont tels que x appartient aux k plus proches voisins de $\mathbf{X}_i$. Des résultats expérimentaux

montrent que les performances de cet estimateur sont souvent meilleures que celles de l'estimateur des k-PPV standard. Un des objectifs de cette thèse sera d'étudier les propriétés (asymptotiques et non-asymptotiques) de cet estimateur, et de savoir dans quelles situations cette variante donne lieu à une amélioration de la vitesse de convergence.

Co-encadrement : Anna Ben-Hamou 50 %, Arnaud Guyader 50 %.

Production des encadrants en lien avec le projet :

Cérou and Guyader [7]
Biau et al. [5]
Biau et al. [4]
Guyader and Hengartner [9]
Biau et al. [6]
Ben-Hamou et al. [1]
Ben-Hamou et al. [2]

Profil de l'étudiant.e recherché.e : Étudiant.e ayant obtenu un Master, ou un diplôme d'une école d'ingénieurs, en aléatoire.

Références

[1] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1) :249–287, 2017. ISSN 1350-7265. doi : 10.3150/15-BEJ743. URL `https://doi.org/10.3150/15-BEJ743`.

[2] A. Ben-Hamou, Y. Peres, and J. Salez. Weighted sampling without replacement. *Braz. J. Probab. Stat.*, 32(3) :657–669, 2018. ISSN 0103-0752. doi : 10.1214/17-BJPS359. URL `https://doi.org/10.1214/17-BJPS359`.

[3] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

[4] G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11(2), 2010.

[5] G. Biau, F. Cérou, and A. Guyader. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory*, 56(4) : 2034–2040, 2010.

[6] G. Biau, F. Cérou, and A. Guyader. New insights into approximate bayesian computation. In *Annales de l'IHP Probabilités et statistiques*, volume 51, pages 376–403, 2015.

[7] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM : Probability and Statistics*, 10 :340–355, 2006.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[9] A. Guyader and N. Hengartner. On the mutual nearest neighbors estimate in regression. *Journal of Machine Learning Research*, 14(8), 2013.

[10] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Avis favorable

Directeur ED386