Sorbonne Université
M2 Mathématiques de la Modélisation
UM5MAM99

Année 2026-2027
Semestre 1

Régression et Classification

1 Sujet

Le but de ce projet est d'étudier certains aspects de l'algorithme du plus proche voisin baggé, tel que décrit dans l'article :

— On the Rate of Convergence of the Bagged Nearest Neighbor Estimate, G. Biau, F. Cérou, and A. Guyader. Journal of Machine Learning Research, JMLR 11, 687-712, 2010.

Plus précisément, on s'intéresse à la version "bagging with replacement", qui correspond à l'estimateur $r_n(\mathbf{x}) = \sum_{i=1}^n V_i \, Y_{(i)}(\mathbf{x})$ où

$$V_i := \left(1 - \frac{i-1}{n}\right)^k - \left(1 - \frac{i}{n}\right)^k .$$

2 Le rapport

Vous pourrez préparer un notebook Python expliquant de manière claire et concise le problème et les simulations effectuées. Les figures devront avoir des titres et légendes suffisamment explicites.

2.1 Partie théorique

Après avoir introduit le contexte et les notations du "bagging with replacement", et compte tenu de ce qui a été présenté en cours, vous montrerez en quoi la borne obtenue pour la vitesse de convergence en Corollaire 2.1 puis en Théorème 2.2 de l'article ci-dessus peut être améliorée dans le cas $d = 2$. Pour ce faire, outre les notes de cours, vous pourrez vous appuyer sur les Théorèmes 2.4 et 14.5 de l'ouvrage *Lectures on the Nearest Neighbor Method*, G. Biau and L. Devroye (Springer, 2015) et adapter la démonstration du Lemme 6.1 de l'article On Deviation Probabilities in Non-parametric Regression with Heavy-Tailed Noise, A. Ben-Hamou and A. Guyader, ALEA, Vol. 22, 775-797, 2025.

2.2 Simulations

Que ce soit en régression ou en classification supervisée, vous pourrez par exemple reprendre les questions traitées en TP et comparer aux bornes théoriques ainsi qu'aux résultats obtenus par la méthode des plus proches voisins.