Données aberrantes et MOM (médiane et moyennes) en apprentissage statistique

Stéphane Gaïffas

13 janvier 2020

Résumé. La grande majorité des méthodes d'apprentissage statistique reposent sur un principe de minimisation du risque empirique, qui amène à un problème d'optimisation de moyenne de fonctions de "perte" associées à chaque échantillon de données, sous l'hypothèse fondamentale que les données sont indépendantes et identiquement distribuées. Par ailleurs, un problème récurrent en science des données est la présence de données aberrantes, ce qui contredit cette hypothèse fondamentale, qui est à la base de la plupart des garanties théoriques disponibles dans la littérature pour les méthodes d'apprentissage statistique.

Des méthodes dites "robustes" ont été développées depuis les années 70 pour rendre les méthodes d'apprentissage moins sensibles aux données aberrantes. Un des principes considérés dans ce cadre est l'utilisation de médiane de moyennes (MOM=Median Of Means) à la place de moyennes de fonctions de perte pour mesurer la qualité de l'attache aux données d'une procédure, et des travaux récents tels que [1] et [2] proposent une étude théorique du principe MOM dans le cadre général de l'apprentissage supervisé.

Le travail proposé dans ce mémoire est donc d'étudier la procédure MOM tant d'un point de vue théorique qu'appliqué, en produisant des expériences illustrant la robustesse de cette approche (notamment en comparaison avec des approches reposant sur une modification des fonctions de perte) et l'étude des garanties théoriques proposées pour MOM.

Références

[1] G. Lecué and M. Lerasle. Robust machine learning by median-of-means : theory and practice. *arXiv :1711.10306 [math, stat]*, Nov. 2017. arXiv : 1711.10306.

[2] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via MOM minimization. *arXiv :1808.03106 [math, stat]*, Aug. 2018. arXiv : 1808.03106.