Initiative de recherche Modèles et traitements mathématiques des données en très grande dimension

Partenaires : GENES - Médiamétrie

Responsable Scientifique : Guillaume Lecué (ENSAE)

Site internet : https://www.institutlouisbachelier.org/programme/modeles-et-traitements-mathematiques-des-donnees-en-tres-grande-dimension/

DESCRIPTION DU PROGRAMME DE RECHERCHE

Dans le cadre de la mesure d'audience (internet, télévision, ...), Médiamétrie collecte et traite des volumétries importantes de données de nature hétérogène. Le volume et la diversité de ces données nécessitent de repenser certains fondamentaux de la statistique tout en proposant de nouvelles idées. Cette évolution est rendue possible par l'avènement de nouvelles technologies informatiques de stockage et de calcul, comme par exemple les architectures MapReduce et Hadoop.

Face aux mégadonnées, les outils mathématiques utilisés relèvent de plusieurs disciplines comme la statistique, l'apprentissage automatique ou l'optimisation. Par ailleurs, ces mégadonnées peuvent provenir de contextes variés : web, téléphone mobile, box ADSL ou autres... Même si les idées et les méthodes sous-jacentes sont souvent similaires, chaque type de données nécessite des algorithmes spécifiques. L'accès plus répandu à ce type de données a permis l'essor des algorithmes de deep learning et pourrait accélérer le développement de nouvelles méthodes.

Dans ce contexte, Médiamétrie souhaite renforcer la R&D sur ces thématiques par des partenariats avec le monde académique.

Les recherches s'orientent autour de 3 thèmes :

1. Rapprochement/fusion de bases de données
2. Enrichissement réciproque de données mixtes et hétérogènes (échantillons vs données exhaustives)
3. Les évolutions du Machine Learning : quelles innovations ?

EQUIPE DE RECHERCHE

- Guillaume Lecué, ENSAE
- Valentin Patilea, ENSAI

PUBLICATIONS DE L'ANNEE

- G. Lecué and Z. Shang **Geometrical viewpoint on the benign overfitting property of the minimum $\ell 2$-norm interpolant estimator**. Submitted 2022.
- J. Depersin and G. Lecué. **Robust subgaussian estimator of a mean vector in nearly**
- **linear time Ann. Statist.** 50(1) : 511-536 (February 2022).