Séminaire de la chaire PARI
Ouvert aux professionnels et aux académiques, accès libre

Séance 18

17 avril 2019

Expliquer ou prédire: données massives et nouveaux défis

Présentation

La plupart des méthodes statistiques classiques (estimation, tests, modèle linéaire etc.) ont été conçues dans un monde de rareté des données. La disponibilité de données massives change radicalement la manière de faire de la statistique : tout devient significatif, les modèles parcimonieux sont rejetés. Le mariage avec le *machine learning* conduit à privilégier la culture de la modélisation prédictive au détriment de celle des modèles génératifs pour reprendre la terminologie de Leo Breiman. C'est en réalité la notion même de modèle qui change : un modèle prédictif n'est qu'un algorithme. La combinaison de plusieurs algorithmes (méthodes d'ensemble) améliore encore leurs performances comme en témoigne le succès des forêts aléatoires, du *boosting* et du *stacking* (encore appelé *blending*). Une autre approche plus ancienne, mais moins connue, consiste à rechercher des modèles locaux.

Opposer prédire à comprendre est classique dans le domaine des données massives : un bon modèle prédictif n'a pas à imiter le mécanisme générateur des données. On sait moins que de tels paradoxes existent aussi pour des *small data*. L'interprétation de modèles « simples » ne va pas non plus de soi; des travaux récents rappellent la difficulté de mesurer l'importance des variables en régression. Comprendre pour mieux prédire: les efforts actuels pour relier *Big Data* et inférence causale permettront sans doute de relever ce défi.

Intervenant : Gilbert Saporta est professeur émérite de statistique appliquée au CNAM

La chaire PARI (programme sur l'appréhension des risques et des incertitudes), portée par l'ENSAE et Sciences Po, a pour objectif d'identifier et comprendre (i) le champ de pertinence de nos outils d'appréhension des risques, et (ii) leurs conditions d'émergence et d'utilisation. Créée début 2015, elle organise un séminaire de recherche mensuel de 2h pour présenter et échanger autour de ses travaux et des thématiques connexes. Le deuxième cycle de la chaire porte sur les enjeux du big data pour l'assurance.