INSTITUT DE FRANCE
Académie des sciences

Cérémonie du 29 mai 2018

Allocution d'Éric Moulines
Apprentissage bayésien

Élu dans la section des Sciences mécaniques et informatiques

Le traitement statistique de l'information est au cœur de mon travail de recherches. Pour le situer dans l'espace de la recherche scientifique, le traitement statistique de l'information est un domaine vaste qui s'est développé aux marges de l'informatique et des mathématiques appliquées. Son importance a explosé au cours des 30 années qui séparent mes premiers pas dans la recherche et cet après-midi.

Les trois applications du traitement de l'information sur lesquelles j'ai été amené à travailler sont, dans l'ordre chronologique des recherches que j'ai menées : le traitement de la parole, le traitement statistique du signal et le traitement des données massives.

Pour la parole, que j'ai abordée à la fin des années 1980, l'enjeu était d'interagir avec les ordinateurs de manière naturelle et pour cela les ordinateurs devaient pouvoir « parler », « comprendre », mais aussi « dialoguer ». Les progrès ont été considérables en 30 ans et les approches statistiques, en reconnaissance de parole comme en synthèse se sont imposées.
Le traitement statistique du signal auquel je me suis ensuite intéressé de façon en dépassant le cas particulier de la parole est une science qui est apparue au milieu du 20ème siècle mais qui a pris son véritable essor au début des années 1970 avec le développement des moyens de calcul et des capteurs. Le traitement statistique du signal a des applications innombrables, et pour le chercheur c'est un terrain de jeu d'une variété sans limite.

Pour finir, les données : c'est aujourd'hui un truisme de dire que la taille des données collectées croît de manière exponentielle. Les sources sont très diverses : des grands instruments de la recherche scientifique aux objets connectés, des médias sociaux aux rayons cosmiques. Le « Big Data » a en quelques années pénétré de très nombreux domaines d'activités où le « graal » reste d'extraire une information pertinente de ces masses de données.

Bien que des avancées remarquables aient été réalisées, dont certaines ont eu un très fort retentissement médiatique, de très nombreuses questions sont encore ouvertes et de nouveaux challenges apparaissent chaque jour.

Le traitement statistique de l'information est un corpus de théories, de méthodes et d'algorithmes dont l'objectif ultime est de combler le fossé entre les données d'un côté et le sens et la décision de l'autre. L'approche que j'ai toujours poursuivie est basée sur l'utilisation de « modèles statistiques bayésiens ».

Le paradigme fondamental des « modèles statistiques » est de considérer que les observations sont une réalisation d'un processus aléatoire gouvernée par une famille de lois de probabilité. L'approche bayésienne fournit un cadre mathématique pour mener un raisonnement plausible en présence d'incertitudes sur les données et des modèles.

On peut très bien traiter l'information en faisant abstraction des modèles : les succès récents des réseaux de neurones et de l'apprentissage profond en sont la preuve éclatante. Mais ce n'est pas la voie que j'ai suivie.

Le mot même de « modèle » implique simplification et idéalisation : la « boite noire » qui a produit les données reste, à de rares exceptions, inconnue. Loin de moi l'idée qu'un système complexe qu'il soit physique, biologique, puisse se résumer à des lois de probabilité, fussent-elles raisonnablement sophistiquées.

Je crois toutefois possible, dans de très nombreuses situations, de construire des représentations certes idéalisées mais qui parviennent à capturer les « caractéristiques statistiques essentielles » des observations dont nous disposons.

La plupart des avancées pour les approches bayésiennes au cours des 30 dernières années sont liées au développement de nouvelles méthodes numériques, permises à la fois par l'accroissement vertigineux des puissances de calcul mais aussi celui des données disponibles.

Ceci a bouleversé le paysage du traitement bayésien de l'information et a permis d'entreprendre le traitement de vastes ensembles de données complexes à l'aide de modèles statistiques très sophistiqués.

Je voudrais maintenant dresser un bref panorama des sujets de recherche sur lesquels je me suis le plus longuement investi.

Je me suis tout particulièrement intéressé à la modélisation des dépendances temporelles qui sont au cœur des problématiques des séries chronologiques et, bien entendu, du traitement de la parole, du signal+. Les modèles sur lesquels j'ai le plus travaillé sont les « chaînes de Markov cachées » qui est une généralisation des modèles d'états linéaires, inventés au tout début des années 1960.

Une chaîne de Markov cachée modélise l'évolution d'un état observé de façon indirecte. L'hypothèse essentielle est que cet état évolue suivant une dynamique markovienne : à chaque instant, la loi de l'état à l'instant suivant ne dépend que de l'état courant. Les chaînes de Markov cachées forment une classe de modèles particulièrement riche et suffisamment flexible pour s'appliquer à des situations aussi variées que l'analyse du génome, la reconnaissance de parole mais aussi le système de positionnement de votre smartphone.
Pour les aspects algorithmiques, je me suis intéressé aux méthodes de Monte Carlo par Chaînes de Markov et à l'optimisation Stochastique.

Les méthodes de Monte Carlo par Chaînes de Markov (MCMC) sont un ensemble d'algorithmes qui permettent de simuler des lois et d'évaluer numériquement des intégrales, quand la dimension de l'espace est grande.

Les méthodes d'optimisation stochastique sont un ensemble d'algorithmes qui couplent l'optimisation et des heuristiques stochastiques comme le « recuit simulé » ou le « gradient stochastique ».

Pour la simulation comme pour l'optimisation, les enjeux actuels sont le passage à l'échelle dans la complexité des modèles et la taille des ensembles de données.

Je souhaiterais conclure sur quelques notes plus personnelles. Mon activité a toujours été à la frontière entre les sciences de l'ingénieur, les statistiques et les probabilités appliquées. J'ai toujours été attiré par les applications et ma recherche se construit en se nourrissant de problèmes pratiques. Mais j'ai aussi pris beaucoup de plaisir à maitriser des outils mathématiques plus formels dont la puissance et la sophistication me fascinent et c'est un bonheur de pouvoir y apporter ma contribution.

J'ai été amené à travailler avec des collègues extraordinaires sans qui mon parcours de recherche n'aurait pas pu être ce qu'il est. Je crois à l'intelligence collective ! J'ai eu la chance de former un grand nombre d'étudiants, dont certains sont restés des collègues proches et des amis : je leur dois aussi beaucoup.

Pour finir, je voudrais remercier ma famille, toujours présente et aimante, et pardonnant mes week-ends souvent studieux. Je ne saurais trop remercier mes parents, tous deux chercheurs et passionnés par leurs métiers, de m'avoir communiqué le virus de la recherche. Comme Obélix, je suis tombé dedans