Journée Statistique / Apprentissage Paris-Saclay

Vendredi 19 Janvier 2018

9h00 Café d'accueil

10h00 **Romain Couillet** *(CentraleSupélec, Université de Paris-Saclay)*
A random matrix approach to bigdata machine learning

> In this talk, we motivate the use of random matrix theory for the analysis of standard and advanced machine learning methods. After a short methodological introduction, we will show that some classical machine learning tools (kernel approaches, neural networks, spectral clustering) exhibit sometimes unexpected behaviors in the large dimensional regime, thereby triggering a complete reconsideration of their usage. These results allow for a better understanding, and thus a possibility to improve, those methods in view of practical applications.

10h50 Pause Café

11h20 **Sophie Donnet** *(INRA)*
Introduction to stochastic block models for ecological and sociological networks

> Modeling relations between entities (individuals, plants, insects...) is a classical question in sociology or ecology. Clustering individuals according to the observed patterns of interactions allows to uncover a latent structure in the data. Stochastic block model (SBM) and Latent Block models (LBM) are popular approaches for grouping the individuals with respect to their interaction profile. These models include latent random variables, making their likelihood intractable. In order to maximise the likelihood, variational versions of the EM algortihms have proved to be flexible tools. In this talk, I will propose an introduction to these models (and some extensions) with motivations issued from sociology and ecology. I will also present standard inference tools ans some theoretical results.

12h10 **Marco Cuturi** *(ENSAE)*
Generative Models and Optimal Transport

> We present in this talk recent advances on the topic of parameter estimation using optimal transport, and discuss possible implementations for "Minimum Kantorovich Estimators". We show why these estimators are currently of great interest in the deep learning community, in which researchers have tried to formulate generative models for images. We will present a few algorithmic solutions to this problem.

13h00 Déjeuner

14h20 **James Ridgway** *(AGRO ParisTech)*
Concentration of tempered posteriors and of their variational approximations

While Bayesian methods are extremely popular in statistics and machine learning, their application to massive datasets is often challenging, when possible at all. Indeed, the classical MCMC algorithms are prohibitively slow when both the model dimension and the sample size are large. Variational Bayesian methods aim at approximating the posterior by a distribution in a tractable family. Thus, MCMC are replaced by an optimization algorithm which is order of magnitude faster. VB methods have been applied in such computationally demanding applications as including collaborative filtering, image and video processing, NLP and text processing... However, despite very nice results in practice, the theoretical properties of these approximations are usually not known. In this talks I will present a general approach to prove concentration of variational approximations of (tempered) posteriors. Our approach also provides a new look on the assumptions usually required to derive concentration of the posterior in Bayesian statistics. I will illustrate the method on Bayesian matrix completion and logistic regression.

15h10 **Bertrand Thirion** *(INRIA)*
Toward a rigorous statistical framework for functional brain mapping

Functional neuroimaging offers a unique view on brain functional organization, which is broadly characterized by two features : the segregation of brain territories into functionally specialized regions, and the integration of these regions into networks of coherent activity. Functional Magnetic Resonance Imaging yields a spatially resolved, yet noisy view of this organization. It also yields useful measurements of brain integrity to compare populations and characterize brain diseases. To extract information from these data, a popular strategy is to rely on supervised classification settings, where signal patterns are used to predict the experimental task performed by the subject during a given experiment, which is a proxy for the cognitive or mental state of this subject. In this talk we will describe how the reliance on large data copora changes the picture : it boosts the generalizability of the results and provides meaningful priors to analyze novel datasets. We will discuss the challenges posed by these analytic approaches, with an emphasis on computational aspects, and current work to improve model identification in this context.

16h00 Pause café

16h30 **Anna Korba** *(Telecom ParisTech)*
Ranking median regression : Learning to order through local consensus

In the present era of personalized customer services and recommender systems, predicting the value taken by a random permutation Σ, describing the preferences of an individual over a set of items $\{1, \ldots, n\}$ say, based on its characteristics modelled as a r.v. X, is an ubiquitous issue. When error is measured by the Kendall distance, this boils down to recovering conditional Kemeny medians of Σ given X from i.i.d. training examples $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$. For this reason, this statistical learning problem is referred to as ranking median regression here. In this talk, I will introduce the probabilistic theory of ranking median regression we propose and then the concept of local medians which enables us to build efficient predictive rules implemented at a local level (in particular k-nearest neighbor and tree-base methods are investigated).

17h20 Fin de la journée