Journée Statistique / Apprentissage Paris-Saclay

19 Janvier 2017

9h00 Café d'accueil

10h00 **Isabelle Guyon** *(Université Paris-Sud, LRI)*
Network Reconstruction : the Contribution of Challenges in Machine Learning

Networks of influence are found at all levels of physical, biological, and societal systems : climate networks, gene networks, neural networks, and social networks are a few examples. These networks are not just descriptive of the "State of Nature", they allow us to make predictions such as forecasting disruptive weather patterns, evaluating the possible effect of a drug, locating the focus of a neural seizure, and predicting the propagation of epidemics. This, in turns, allows us to device adequate interventions or change in policies to obtain desired outcomes : evacuate people before a region is hit by a hurricane, administer treatment, vaccinate, etc. But knowing the network structure is a prerequisite, and this structure may be very hard and costly to obtain with traditional means. For example, the medical community relies on clinical trials, which cost millions of dollars ; the neuroscience community engages in connection tracing with election microscopy, which take years before establishing the connectivity of 100 neurons (the brain contains billions).
This presentation will review recent progresses that have been made in network reconstruction methods based solely on observational data. Great advances have been recently made using machine learning. We will analyze the results of several challenges we organized, which point us to new simple and practical methodologies.

La reconstruction du réseau : la contribution des défis en Machine Learning

Les réseaux d'influence se retrouvent à tous les niveaux des systèmes physiques, biologiques et sociaux : réseaux climatiques, réseaux de gènes, réseaux de neurones, et réseaux sociaux en sont quelques exemples. Ces réseaux ne sont pas seulement descriptif d'un «état de la nature», ils nous permettent de faire des prédictions telles que la prévision des conditions météorologiques, l'évaluation de l'effet possible d'un médicament, la localisation d'un foyer épileptique, et la prévision de la propagation des épidémies. Ceci nous permet alors d'intervenir pour obtenir les résultats souhaités : évacuer les populations d'une région avant qu'elle soit frappée par un ouragan, administrer un traitement, vacciner, etc. Mais la connaissance de la structure du réseau est une condition préalable à l'application de ces méthodes, et cette structure peut être très difficile et coûteuse à obtenir avec des moyens traditionnels. Par exemple, la communauté médicale s'appuie sur les essais cliniques, qui coûtent des millions de dollars ; la communauté des neurosciences analyse des images de microscopie électorale, ce qui prendra des années avant d'établir la connectivité de 100 neurones (alors que le cerveau en contient des milliards).
Cette présentation examinera les récents progrès qui ont été faits dans les méthodes de reconstruction de réseaux fondées uniquement sur des données d'observation. De grands progrès ont été récemment réalisés grâce à l'apprentissage des machines (machine learning). Nous allons analyser les résultats de plusieurs défis que nous avons organisées, qui nous pointent vers de nouvelles méthodes simples et pratiques.

10h50 Pause Café

11h20 **Matthieu Lerasle** *(CNRS, LMO)*
Learning from MOM's principles

TBA

12h10 **Frédéric Chazal** *(INRIA Saclay)*
Persistent homology for data : stability properties and statistical aspects

Computational topology has recently seen an important development toward data analysis, giving birth to Topological Data Analysis. Persistent homology appears as a fundamental tool in this field. It is usually computed from filtrations built on top of data sets sampled from some unknown (metric) space, providing "topological signatures" revealing the structure of the underlying space. When the size of the sample is large, direct computation of persistent homology often suffers two issues. First, it becomes prohibitive due to the combinatorial size of the considered filtrations and, second, it appears to be very sensitive to noise and outliers. The goal of the talk is twofold. First, we will briefly introduce the notion of persistent homology and show how it can be used to infer relevant topological information from metric data through stability properties. Second, we will present a method to overcome the above mentioned computational and noise issues by computing persistent diagrams from several subsamples and combining them in order to efficiently infer robust and relevant topological information.

13h00 Déjeuner

14h30 **Anne Sabourin** *(Télécom ParisTech)*
Feature clustering for extreme events analysis, with application to extreme stream-flow data

The dependence structure of multivariate extreme events of multivariate nature is a major concern for risk management. In a high dimensional context ($d > 50$), dimension reduction is a natural first step. However, analyzing the tails of a dataset requires specific approaches that standard algorithms such as PCA do not accommodate. One convenient characterization of extremal dependence is the *angular measure*, defined on the positive orthant of the $d-1$ dimensional hyper-sphere, which provides direct information about the probable 'directions' of extremes. Recent works (Goix et al., 2015, 2016) have defined sparsity in multivariate extremes as the concentration of the angular measure on low dimensional subspheres, and have proposed an algorithm detecting such a pattern in cases where the latter is apparent.
Given a dataset that exhibits no clear sparsity pattern, we propose an alternative clustering algorithm allowing to group together the features of the dataset that are 'dependent at extreme level', i.e. that are likely to take extreme values simultaneously. To bypass the computational issues that arise when it comes to dealing with possibly $O(2^d)$ groups of features, our algorithm exploits the graphical structure stemming from the definition of the clusters, which reduces drastically the number of subgroups on which the estimation procedure has to be performed. Results on simulated and real data show that our method allows to recover a meaningful summary of the dependence structure of extremes in a reasonable amount of computational time.

15h20 **Alexandre Allauzen** *(Université Paris-Sud, LIMSI)*
Neural Networks for Natural Language Processing

In the last decade, artificial neural networks have thoroughly renewed the research in Natural Language Processing (NLP). Most of NLP tasks require to model structured data that are characterized by very peculiar and sparse distributions of events along with a large set of possible outcomes. To tackle such challenges in terms of machine learning, neural networks introduced distributed representations of words (or other linguistic units). This kind of approach jointly learns the representation along with the decision process, and opens wide research perspectives. The presentation will review some recent developments in this field. The language modeling and machine translation tasks will serve as leading examples to illustrate the research challenges and to then discuss future directions.

Réseaux de neurones artificiels pour le traitement automatique des langues

Ces dernières décennies, les réseaux de neurones artificiels ont profondément renouvelé les perspectives de recherche en traitement automatique des langues (TAL). La plupart des applications en TAL nécessitent de modéliser des données structurées qui se caractérisent par des distributions particulières, parcimonieuses et avec des espaces de réalisations très grands. Dans ce contexte, les réseaux de neurones ont permis des avancées importantes en introduisant la notion de représentations continues pour le TAL. Néanmoins, d'un point de vue apprentissage automatique, de nombreux défis restent ouverts et sont liés aux spécificités des données langagières. Cet exposé présentera ces défis, les solutions qui sont explorés aujourd'hui, ainsi que les perspectives de recherche pour l'utilisation des modèles neuronaux en TAL.

16h10 Pause café

16h40 **Vianney Perchet** *(CMLA)*
New perspectives for multi-armed bandits and their applications

TBA

17h30 Fin de la journée