Programme Long - YSP 6

Session 1 (9h45- 11h) : Percolation

Tutoriel : Aurélia Deshayes (Paris Diderot)

La percolation a été introduite par Hammersley en 1957 pour comprendre la répartition d'un fluide dans un matériau poreux. On modélise le matériau par un graphe G et sa porosité en gardant chaque site (ou arête) avec une probabilité p. On s'intéresse aux propriétés du graphe aléatoire obtenu en fonction du paramètre p. Une version continue de ce modèle a ensuite été introduite en remplaçant G par le graphe Rd dont on ne garde qu'une sous partie aléatoire constituée de boules (en se donnant un processus de poisson pour leurs centres et une loi de rayons). On s'intéresse à nouveau aux propriétés topologiques du sous espace obtenu. Enfin, de retour dans le monde discret, nous parlerons de la bootstrap percolation qui traite d'une classe d'automates cellulaires dont la condition initiale est une configuration de percolation; on s'intéresse dans ce cas à l'objet obtenu à la fin de la dynamique. La percolation et ses extensions sont des modèles simples à expliquer mais autour desquels il reste encore aujourd'hui de nombreuses questions ouvertes qui mettent aux défis nos meilleurs probabilistes.

Exposé 1 : Florestan Labey (Université François Rabelais - Tours)

Boules carrées, percolation, paramètre critique en grande dimension.

On exhibe dans le cas simple de la norme infinie dans $\mathbf{R}^d$ un équivalent du paramètre critique de percolation booléenne lorsque la dimension tend vers l'infini, et on présente le résultat dans le cas plus général des boules de norme p.

Exposé 2 : Assaf Shapira (Paris Diderot)

Criticalité de la Percolation bootstrap sur des arbres

La percolation bootstrap est une dynamique déterministe en temps discret définie sur un graphe. Chaque sommet du graphe peut être soit sain soit infecté, et à chaque étape de temps un sommet sain devient infecté, si sa voisinage satisfait une certaine condition. On étudiera des modèles de percolation bootstrap sur les arbre réguliers et les arbres de Galton-Watson, en se concentrant sur la quantité de sommets infecté et particulièrement sur sa discontinuité quand les paramètres du modèle changent.

Session 2 (11h30- 12h45) : Optimisation

Tutoriel : Aymeric Dieuleveut (EPFL)
Stochastic algorithms for machine learning

This talk presents an overview of some stochastic algorithms used in large scale machine learning, especially first order methods. We describe the most useful methods, together with classical convergence rates, for the minimization of the empirical risk then for the generalisation risk. In the last part, we focus on constant step size stochastic gradient descent, which is analysed as a Markov chain.

Exposé 1 : Mathurin Massias (Telecom ParisTech & INRIA Saclay)
Generalized Concomitant Lasso for sparse multimodal regression

In high dimension, it is customary to consider Lasso-type estimators to enforce sparsity. For standard Lasso theory to hold, the regularization parameter should be proportional to the noise level, which is generally unknown in practice. A remedy is to consider estimators, such as the Concomitant Lasso, which jointly optimize over the regression coefficients and the noise level. However, when data from different sources are pooled to increase sample size, or when dealing with multimodal data, noise levels differ and new dedicated estimators are needed.We provide new statistical and computational solutions to perform heteroscedastic regression, with an emphasis on functional brain imaging with magneto- and electroencephalographic (M/EEG) signals. When instantiated to de-correlated noise, our framework leads to an efficient algorithm whose computational cost is not higher than for the Lasso, but addresses more complex noise structures. Experiments demonstrate improved prediction and support identification with correct estimation of noise levels. Results on multimodal neuroimaging problems with M/EEG data are also reported.

Exposé 2 : Aude Genevay (Paris Dauphine & ENS Ulm)
Large Scale Optimal Transport

Optimal Transport (OT) is a powerful tool to compare probability distributions, but suffers from two main drawbacks: a heavy computational cost and poor sample complexity. Adding an entropic regularization to the problem has proven very efficient to improve both aspects. After a quick review of the basics of OT, I will present efficient solvers for large scale optimal transport, which are based on stochastic optimization techniques. Eventually, I will use regularized OT for inference in generative models, a type of probabilistic models which don’t have a density and thus can’t be learnt with maximum likelihood.

Session 3 (14h15- 15h30) : Inférence de réseaux

Tutoriel : Marco Corneli (Université Nice Sophia Antipolis)

Graphs and statistical inference, two examples

The tutorial consists of two parts. The first part focuses on gene co-expression networks and details a graphical model (Gaussian concentration) in which the (weighted) adjacency matrix is not observed and must be estimated, accounting for sparsity. Since the number of the model parameters is high compared with the number of observations, a naive inference procedure presents some drawbacks which are discussed. An alternative penalized likelihood method is detailed. The second part of the tutorial focuses on random graphs and an observed adjacency matrix is assumed to be generated according to the stochastic block model (SBM). In SBM, the probability of observing an edge between two nodes only depends on their clusters. Due to the graphical structure of the model, the standard EM algorithm for mixture models cannot be adopted for inference. To tackle this issue a variational approximation based method (V-EM) is illustrated.

Exposé 1: Alyssa Imbert (INRA)

The purpose of this talk is to see how can we increase the reliability of network inference when some samples could not be observed in RNAseq expression data (count data). We will review a graphical model adapted to count data: the log-linear Poisson graphical model (llgm). Then, we present a new method, hot-deck multiple imputation (hd-MI), which is based on imputation for samples without available RNAseq data that are considered as missing data but are observed on a auxiliary dataset (Rt-qPCR, microarray,...). This auxiliary dataset provide external information on gene expression similarity between samples.

Exposé 2: Timothée Tabouy (AgroParistech)

The purpose of this talk is to deal with non-observed dyads during the sampling of a network and consecutive issues in the Stochastic Block Model (SBM) inference. We will review sampling designs and recover Missing At Random (MAR) and Not Missing At Random (NMAR) conditions for SBM. We will also introduce several variants of the variational EM (VEM) algorithm for inferring the SBM under various sampling designs (MAR and NMAR). The sampling design must be taken into account only in the NMAR case. Model selection criteria based on Integrated Classification Likelihood (ICL) are derived for selecting both the number of blocks and the sampling design.

Session 4 (16h- 17h15) : Machine learning et réseaux de neurones

Tutoriel : Warith Harchaoui (Université Paris Descartes - Oscaro.com)
Neural Networks: an Introduction

Exposé 1 : Peter Naylor (Institut Curie - Mines ParisTech)
Deep Neural networks for segmentation. Application to Histopathology

L'analyse et la quantification des données histopathologique est l'outil de référence dans l'élaboration du diagnostic des individus souffrant du cancer. En particulier, le médecin regarde les populations cellulaires sur ces images. Nous proposons une solution pour repérer de manière automatique et précise les cellules grâce à des méthodes de deep learning pour la segmentation.

Exposé 2 : Stanislas Chambon (Rythm - Telecom ParisTech)
Deep learning for prediction on EEG

L'exposé portera sur la classification de stades de sommeil à partir d'enregistrements de polysomnographie (électroencéphalogramme, électro-oculogramme… au cours d'une nuit de sommeil) qui se présentent sous la forme d'une série temporelle multivariée. La méthode présentée se base sur réseau de neurones convolutionnel entraîné directement sur des séries temporelles multivariées, pour classifier les stades de sommeil d'un sujet. L'influence de différents facteurs sera présentée et analysée.