Programme détaillé

9h30 - 10h00 ***Fairness and explainability of algorithms: definitions, paradoxes and biases*** par Gilbert Saporta (CNAM, CEDRIC)

Résumé : Fairness of algorithms is the subject of a large body of literature, guides, computer codes and tools. Machine Learning and AI algorithms commonly used to accept loan applications, select responses to job offers, etc. are often accused of discriminating against groups. We will begin by examining the relationship between fairness, explainability, and interpretability. One might think that it is better to understand how an algorithm works in order to know whether it is fair, but in fact this is not the case, because transparency or explainability are relative to the algorithm, whereas fairness concerns its differential application to groups of individuals. There is a wide variety of often incompatible measures of fairness. Moreover, questions of robustness and precision are often ignored. The choice of a measure is not only a matter of statistical considerations, but of ethical choices. The "biases" of the algorithms are often only the reproduction of those of previous decisions found in the training data. But they are not the only ones. We will attempt to draw up a typology of the main biases: statistical, societal, cognitive, etc. and to discuss the links with causal models.

10h15 - 11h15 ***Identifying early help referrals for local authorities with machine learning and bias analysis*** par Eufrasio de A. Lima Neto (School of Computer Science and Informatics, De Montfort University)

Résumé : Local authorities in England, such as Leicestershire County Council (LCC), provide Early Help services that can be offered at any point in a young person's life when they experience difficulties that cannot be supported by universal services alone, such as schools. This paper investigates the utilisation of machine learning (ML) to assist experts in identifying families that may need to be referred for Early Help assessment and support. LCC provided an anonymised dataset comprising 14 360 records of young people under the age of 18. The dataset was pre-processed, ML models were developed, and experiments were conducted to validate and test the performance of the models. Bias-mitigation techniques were applied to improve the fairness of these models. During testing, while the models demonstrated the capability to identify young people requiring intervention or early help, they also produced a significant number of false positives, especially when constructed with imbalanced data, incorrectly identifying individuals who most likely did not need an Early Help referral. This paper empirically explores the suitability of data-driven ML models for identifying young people who may require Early Help services and discusses their appropriateness and limitations for this task.

11h15 - 12h30 ***On Selection Bias and Fairness Issues in Machine Learning*** par Stéphane Clémençon (Telecom Paris, LTCI)

Résumé : With the deluge of digitized information in the Big Data era, massive datasets are becoming increasingly available for learning predictive models. However, in many situations, the poor control of the data acquisition processes may jeopardize the outputs of machine-learning algorithms and selection bias issues are now the subject of much attention. Recently, the accuracy of facial recognition algorithms for biometrics applications has been fiercely discussed for instance, its monitoring over time revealing sometimes a predictive performance very far from what was expected at the end of the

training stage. The use of machine-learning methods for designing medical diagnosis/prognosis support tools is currently triggering the same type of fear. Making the enthusiasm and the confidence for what can be accomplished by machine learning durable requires to revisit practice and theory both at the same time. It is precisely the purpose of this talk to explain and illustrate through real examples how to extend Empirical Risk Minimization, the main paradigm of statistical learning, when the training observations are biased, i.e. are drawn from distributions that may significantly differ from that of the data in the test/prediction stage. As expected, there is 'no free lunch': practical, theoretically grounded, solutions do exist in a variety of contexts (e.g. training examples composed of censored/truncated/survey data) but their implementation crucially depends on the availability of relevant auxiliary information about the data acquisition process. One should also have in mind that the 'bias' in machine-learning, as perceived by the general public, also refers to situations where the predictive error exhibits a huge disparity, to cases where the predictive algorithms are much less accurate for certain population segments than for others. If certain facial recognition algorithms make more mistakes for certain ethnic groups for instance, representativeness issues concerning the training data should not be incriminated solely: the variability in the error rates can be due just as much to the intrinsic difficulty of certain recognition problems or to the limitations of the state-of-the-art machine-learning technologies. As will be discussed in this talk, trade-offs between fairness and predictive accuracy then become unavoidable.