Talk Abstracts

0. Keynote Address by Inderpal Bhandari

Inderpal Bhandari, Global Chief Data Officer of IBM, will share IBM's internal experience in accelerating enterprise data & AI. He will discuss how to "Develop a clear data strategy, Attributes of AI, AI Enterprise Blueprint, and how to turn data into business value".

1. Machine Learning in a Nutshell by Nikhil R. Pal

Abstract: This talk will be quite different from the rest of the talks in the Workshop. I shall not focus on any specific topic and provide a detailed exposition on that. I shall begin with a brief discussion on three related (often confusing to me) terms: Data Science, Machine learning, and Artificial Intelligence. This may convince us that Machine Learning (ML) is at the core of all three. Then I shall discuss three primary families of problems (clustering, classification, and feature analysis) that ML usually addresses, and issues/difficulties related to them – all at the conceptual level. This will be followed by a brief discussion on various learning paradigms (supervised, unsupervised, semi-supervised, reinforcement, and transfer learning) and the type of problems they can solve. I plan to conclude with a few (I have a big list though) important issues that in my view need more attention while designing ML systems.

2. Regression and Prediction by **Probal Chaudhuri**

Abstract: The lecture will begin with the introduction of linear least squares regression along with a brief historical review. Computational issues especially for high dimensional data and some relevant statistical facts will be discussed. Then the connection between regression and prediction problems will be described. Some important and widely used alternatives to linear regression will be introduced, and their implementation in data will be discussed. These will include weighted least squares method in nonlinear and generalized regression models. Towards the end, I shall introduce nonparametric regression and highlight its applications.

3. Automatic model building in mixed effect models: applications in population pharmacology by **Marc Lavielle**

Abstract: Building and validating a mixed effects model are generally difficult and laborious tasks for the modeler. Indeed, it requires to find the "best" covariate model, i.e. to identify which covariates significantly explain the variability of some individual parameters, to identify the "best" correlation model for the random effects, and to find the "best" residual error model for continuous data. I will present an extension of the EM algorithm that allows one to automatically build a linear mixed-effects model by optimizing a penalized likelihood criterion iteratively. I will also present the SAMBA (Stochastic Approximation for Model uilding Algorithm) algorithm, an extension of this method for nonlinear mixed effects models. Once the model is built, it must be validated, i.e. each of the hypotheses made on the model must be tested (covariate model, correlation structure of the random effects, distribution of the random effects, distribution of residual errors, etc.). I will show how to construct unbiased hypothesis tests to validate each of these hypotheses. These methods for building and validating mixed effects models are implemented in Monolix and in the Rsmlx package (http://rsmlx.webpopix.org).

4. Optimal transport for machine learning by Lénaïc Chizat

Abstract: Optimal transport theory originated from the problem of finding the most efficient way of moving a pile of dirt from one configuration to another. Once put in a mathematical form, it provides extremely useful tools for comparing, interpolating and processing objects such as distributions of mass, probability measures or histograms. In this talk, we will review the basics of optimal transport theory, covering statistical and algorithmic aspects, and its use in machine learning.

5. Sensitivity Analysis from computer code experiment to statistical learning by Fabrice Gamboa

Abstract: Computer code experiment deals with statistical techniques in order to supervise and interpret numerical results produced by computer and/or physical experiments. Roughly speaking, one deals with a non-parametric relation Y=F(X). Where Y is the output of the system (the result given by the computer or the physical experience), and X is a random vector modeling the system inputs describing the experience scenario. One of the main question occurring in this scientific paradigm is the ranking of the influence of each component of X on Y. This problem is called sensitivity analysis. We will give an overview of various statistical techniques to perform sensitivity analysis and will also discuss how these techniques are useful in statistical learning.

6. Zero learning at Facebook by **Olivier Teytaud**

Abstract: We present recent research trends in (possibly partially observable) Markov Decision Processes, combining Monte Carlo Tree Search and Deep Learning with zero learning. In particular we made the first wins against humans in challenging connection games such as Hex and Havannah.

7. Fair Learning by Jean-Michel Loubes

Abstract: We consider the issue of bias in machine learning, aka fair learning. In some cases, the learning sample may present some bias that could possibly be learnt by the algorithm and then propagated to the entire population through automatic decisions, providing a mathematical legitimacy for this unfair treatment. When giving algorithms the power to make automatic decisions, the danger may come that the reality may be shaped according to their prediction, thus reinforcing their beliefs in the model which is learnt. Hence, achieving fair treatment is one of the growing fields of interest in machine learning.

In the fair classification setup, we recast the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter. We propose a Random Repair which yields a trade-off between minimal information loss and a certain amount of fairness.

8. Threats and Security Risks in Data Science by Saibal K. Pal

Abstract: Today, data is considered to be one of the world's most valuable resources. A number of important applications demand collection, storage, transmission, processing and analysis of digital data, some of which may be sensitive and not suitable for public consumption. In the wake of frequent data breaches, cyber-attacks, misinterpretation & misuse by global corporations and collection by nation states for surveillance, there is an urgent requirement for strict regulations as well as practical solutions for its protection and proper use.

Also, the last decade has witnessed rapid progress in Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) technologies and their applications. Self-learning and self-enhancement capabilities of the current generation of AI systems have led to their steady improvement in performance for many applications. In addition to the socially useful applications, ML systems are now being developed for automated cyber-attacks and military warfare. Algorithmic Decision Systems (ADS) are also being used in many critical medical, legal and financial applications. However, a number of serious issues including algorithmic bias, malfunctioning and failures of ADS have raised technical, ethical, political and legal concerns.

Common attack vectors and defensive solutions for ADS will be discussed. Different security issues in Data Science right from planning & development phase to systems deployment would be presented along with technical solutions. Potential areas of research for their secure design, development and deployment would also be highlighted.

9. AI for smarter cities in India by Raghu Krishnapuram

Abstract: Most of the efforts around the world to build applications of data science and artificial intelligence for smarter cities have been focussed on scenarios in developed countries. In emerging markets such as India, many of the assumptions on the nature and quality of data, as well as affordability and availability of infrastructure, are not valid. As a result, we need to approach the problems differently. This talk provides a brief overview of an architecture (developed at the Robert Bosch Centre for Cyber-Physical Systems, IISc) for enabling data-driven applications for smarter cities in India, and describes in some detail specific data science and AI applications of that are illustrative of the peculiar challenges that need to be addressed.

10. Machine Learning at Amazon by Vineet Chaoji

Abstract: The talk will provide an overview of the variety of ML problems encountered, throughout the customer lifecycle, at a large e-commerce company such as Amazon. Further, the talk will pick out a few specific problems and go into further details around the ML modeling and results. The talk will conclude with potential challenges and areas where academia and industry can collaborate.

11. Data Science for Connected and Autonomous Vehicles by **Hillol Kargupta**

Abstract: Data science is playing an increasingly important role in connected and autonomous vehicle technology. Vehicle health modeling, driver scoring, fuel cost optimization, and autonomous navigation are just some of the many areas where data-driven technology is disrupting the market, impacting not only our experience with vehicles, but almost all aspects of our life. This talk will focus on the technology for the current and next generation of connected and autonomous vehicles. It will start by identifying a few applications in the areas of vehicle parts inventory management, insurance claims processing, health-care and location-based services that call for distributed edge analytics onboard the vehicle and also in-cloud data analysis. Next, it will offer a perspective on how these applications can be addressed by adapting the current generation of connected vehicle products that are scaling rapidly worldwide. The talk will also discuss analytical foundations for developing data science algorithms geared towards such applications. It will end with a historical perspective of the speaker's commercial and academic experience in building connected vehicle products and some of the related research areas that are emerging in the horizon.

12. Learning with signatures by **Gérard Biau**

Abstract: Sequential and temporal data arise in many fields of research, such as quantitative finance, medicine, or computer vision. In this talk I will discuss a novel approach for sequential learning, rooted in rough path theory. Its basic principle is to represent multidimensional paths by a graded feature set of their iterated integrals, called the signature. This approach relies critically on an embedding principle, which consists in representing discretely sampled data as paths. After a survey of machine learning methodologies for signatures, I will investigate the influence of embeddings on prediction accuracy with an in-depth study of three recent and challenging datasets. With a good embedding, the signature combined with a simple algorithm achieves results competitive with state-of-the-art, domain-specific approaches.

13. Application of AI in Consumer Centric Business by **Pratik Pal**

Abstract: to be updated soon.

14. Stochastic algorithms for entropic optimal transport by **Bernard Bercu**

Abstract: The statistical analysis of high-dimensional data using tools from optimal transport theory has recently gained increasing popularity. This talk is devoted to Robbins-Monro stochastic algorithms in order to estimate the entropically regularized Wasserstein distances between two probability measures. Our main contribution is to establish the almost sure convergence and the asymptotic normality of our estimates in the discrete and semi-discrete settings. Numerical experiments on the optimal mapping between the distribution of spatial locations of reported incidents of crime in Chicago and the locations of Police stations are also provided to illustrate the usefulness of our approach. This is a joint work with Jeremie Bigot.