# Optimization in Machine Learning and Applications

*Francesco Archetti and Antonio Candelieri*

Department of Computer Science, University of Milano-Bicocca

Machine Learning and Optimization are intertwined at several critical junctions. Still, classical optimization underpins a different reference model, basically deterministic, while Machine Learning deals with “bounded rationality” and must work with partial and uncertain information. This situation poses new challenges for optimization algorithms which must “explore” the design space to acquire – at a cost – new information and “exploit” what is available to approximate the optimal design.

To make this point this talk aims at the problem of Automatic Algorithm Configuration and hyperparameter optimization, which is general in Machine Learning, in the specific case of predictive analytics and shows its solution requires stochastic global optimization.

Among the plethora of methods proposed so far, the talk focuses on Bayesian Optimization which offers a principled way to handle the trade-off between exploration and exploitation, a model flexible enough to accommodate different design spaces and gained wide acceptance in academia and industry.

Such a flexibility allows also for addressing complex real-life problems, more precisely those related to optimization of industrial systems where constraints – in many cases unknown a-priori – might significantly shrink the feasible region within the design space. In some case these constraints directly come from the aim to avoid the evaluation of designs which can result in a disruption of system to be optimized (i.e. constrained and safe Bayesian Optimization). Finally, Approximate Dynamic Programming – aka Reinforcement Learning – will be briefly presented as another solution for these applications, based on the paradigm “learn and optimize” (i.e. defining the next decision to know something more about the system while trying to optimize it according to the knowledge acquired so far).

# Mathematics of Deep Learning

*Alexander Balinsky*

School of Mathematics, Cardiff University

Deep Learning is another name for a set of algorithms that use a neural network as an architecture. In the past few years, Deep Learning has generated much excitement due to many breakthrough results in speech recognition, computer vision and text processing. This recent success has been due to new mathematical techniques, the availability of inexpensive, parallel hardware (GPUs, computer clusters) and massive amounts of data. This powerful way of processing data can be used to address an ever-growing number of problems, and its impact on science and society is increasing exponentially. In this talk we present mathematical foundations of Deep Learning, relations with statistical physics and features extraction.

# Reliable Machine Learning by Conformal Predictors

*Alexander Gammerman*

Centre for Machine Learning (CLRC), Royal Holloway and Bedford New College, University of London

The talk reviews a modern machine learning technique called Conformal Predictors. The approach has been motivated by algorithmic notion of randomness and allows us to make reliable predictions with valid measures of confidence for individual examples. The developed technique guarantees that the overall accuracy can be controlled by a required confidence level. Unlike many conventional techniques the approach does not make any additional assumption about the data beyond the iid assumption: the examples are independent and identically distributed. The way to test this assumption is described. The talk also illustrates applications of Conformal Predictors and their various extensions like Inductive conformal predictors, Mondrian machines, ridge regression confidence machine and probabilistic predictors called Venn machines, to many different fields including medicine, cheminformatics, information security, environment, plasma physics, home security and others.

# Introduction to Conformal Prediction

*Lars Carlsson*

Stena Line, Gothenburg, Sweden and Visiting Professor at Royal Holloway, University of London

How good is your prediction? In risk-sensitive applications, it is crucial to be able to assess the quality of a prediction, however, traditional classification and regression models don't provide their users with any information regarding prediction trustworthiness. In contrast, conformal classification and regression models associate each of their multi-valued predictions with a measure of statistically valid confidence, and let their users specify a maximal threshold of the model's error rate - the price to be paid is that predictions made with a higher confidence cover a larger area of the possible output space. This tutorial aims to provide its attendees with the knowledge necessary to implement conformal prediction in their daily data science work, be it research or practice oriented, as well as highlight current research topics on the subject.

Since its development the framework has been combined with many popular techniques, such as Support Vector Machines, k-Nearest Neighbours, Neural Networks, Ridge Regression etc., and has been successfully applied to many challenging real world problems, such as the early detection of ovarian cancer, the classification of leukaemia subtypes, the diagnosis of acute abdominal pain, the assessment of stroke risk, the recognition of hypoxia in electroencephalograms (EEGs), the prediction of plant promoters, the prediction of network traffic demand, the estimation of effort for software projects and the back calculation of non-linear pavement layer moduli. The framework has also been extended to additional problem settings such as semi-supervised learning, anomaly detection, feature selection, outlier detection, change detection in streams and active learning. The aim of this symposium is to serve as a forum for the presentation of new and ongoing work and the exchange of ideas between researchers on any aspect of Conformal Prediction and its applications.

# Stochastic variants of classical optimization methods, with complexity guarantees

*Coralia Cartis*

Oxford University

Optimisation is a key component of machine learning application, as it helps with training of (neural net, nonconvex) models and parameter tuning. Classical optimization methods are challenged by the scale of ML applications and the lack of /cost of full derivatives, as well as the stochastic nature of the problem. On the other hand, the simple approaches that the ML community uses need improvement. Here we try to merge the two perspectives and adapt the strength of classical optimization techniques to meet the challenges of ML applications.

# Visual Analytics for High-Dimensional Data Exploration and Engineering Design Optimisation

*Timoleon Kipouros*

Department of Engineering, University of Cambridge

A dataset with M items has 2^M subsets anyone of which may be the one satisfying our objective. With a good data display and interactivity, our fantastic pattern-recognition defeats this combinatorial explosion by extracting insights from the visual patterns. This is the core reason for data visualisation. With parallel coordinates, the search for relations in multivariate data is transformed into a 2-D pattern recognition problem. This is particularly important in modern industries where the design and optimisation in complex multidisciplinary design spaces is crucial. It is a process involving the simultaneous consideration of conflicting multiple criteria stemming from different domains and stakeholders. Computational models and simulations are utilised extensively, and multidimensional visualisation, as will be shown, can play a key role exploiting our fantastic pattern recognition ability in discovering relational information in such datasets and sequentially guiding the complex engineering design decisions. This will happen by identifying the connections between different stakeholder expectations and engineering technical properties satisfying a number of physical, geometrical, and any type of constraints. Though the examples will be from the aerospace industry, the methodology is widely applicable.

# The Maximum Mean Discrepancy for Training Generative Adversarial Networks

*Arthur Gretton*

Gatsby Computational Neuroscience Unit, University College London

Generative adversarial networks (GANs) use neural networks as generative models, creating realistic samples that mimic real-life reference samples (for instance, images of faces, bedrooms, and more). These networks require an adaptive critic function while training, to teach the networks how to move improve their samples to better match the reference data. I will describe a kernel divergence measure, the maximum mean discrepancy, which represents one such critic function. With gradient regularisation, the MMD is used to obtain current state-of-the art performance on challenging image generation tasks, including 160 x 160 CelebA and 64 x 64 ImageNet. In addition to adversarial network training, I'll discuss issues of gradient bias for GANs based on integral probability metrics, and mechanisms for benchmarking GAN performance. Time permitting, I'll also briefly cover hypothesis testing applications of MMD.

# Peculiarities of High Dimensions

*Anatoly Zhigljavsky*

School of Mathematics, Cardiff University

In this talk, we discuss several simple features of high-dimensional bodies and argue that human intuition could be completely wrong about these bodies. We also discuss some counter-intuitive properties of probability distributions in high-dimensional spaces. From our discussion, we will draw some conclusions about unwelcoming properties of some well-known optimization and data analysis algorithms.

# Multi-word Expressions in Text Mining

*Irena Spasic*

School of Computer Science, Cardiff University

The increasing amount of textual information in requires effective term recognition methods to identify textual representations of domain-specific concepts as the first step toward automating its semantic interpretation. The dictionary look-up approaches may not always be suitable for dynamic domains such as biomedicine or the newly emerging types of media such as patient blogs, the main obstacles being the use of non-standardised terminology and high degree of term variation. Term conflation is the process of linking together different variants of the same term. In automatic term recognition approaches, all term variants should be aggregated into a single normalized term representative, which is associated with a single domain-specific concept as a latent variable. FlexiTerm is an unsupervised method for recognition of multi-word terms from a domain-specific corpus. It uses regular expressions to constrain the search space based on term formation patterns and then processes them statistically to identify largest frequently occurring bags of words and the corresponding terms.

FlexiTerm uses a range of methods to normalize three types of term variation - orthographic, morphological, and syntactic variations.

Acronyms, which represent a highly productive type of term variation, were not originally supported. In this talk, we describe how the functionality of FlexiTerm has been extended to recognize acronyms and incorporate them into the term conflation process. We evaluated the effects of term conflation in the context of information retrieval as one of its most prominent applications. On average, relative recall increased by 32 points, whereas index compression factor increased by 7% points. Therefore, evidence suggests that integration of acronyms provides non-trivial improvement of term conflation.

http://users.cs.cf.ac.uk/I.Spasic/flexiterm/