Machine learning

Centre International de Mathématiques et d'Informatique
TOULOUSE

Sept.-Dec.2015
Trimestre Thématique
Machine Learning

Workshop 3
Learning with Structured Data And Applications to Natural Language and Biology

Dec. 9-10

Deep Learning Tutorial

Dec. 11

INSTITUT de MATHEMATIQUES de TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER

IRIT
CNRS - INPT - UPS - UT1 - UT2J

Workshop 3: Learning with Structured Data and applications on Natural Language and Biology

December 9-11, 2015

This workshop aims at presenting recent advances in Machine Learning field of Structured Prediction. The goal of structured prediction is to learn from data for which an underlying structure exists (e.g. a graph) as well as produce a graph as output. Both theoretical and practical issues will be addressed by the invited speakers. Topics include structured and incremental perceptrons, Maximum Entropy Markov Models, Conditional Random Fields, Maximum Margin Markov Networks, SVMs for Interdependent and Structured Outputs. particular attention will be given to applications in Natural Language Processing (in areas such as syntactic analysis and discourse analysis) and biology.

The last day of the workshop will be devoted to a special course "deep learning : background and application to natural language processing" given by Alexandre Allauzen (LIMSI, UniversitÃĪ Paris-Sud XI).

List of invited speakers

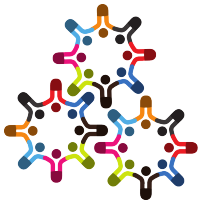
- Alexandre Allauzen, University Paris-Sud XI
- Eustasio del Barrio, University of Valladolid
- Xavier Carreras, Xerox
- Francois Coste, Inria Rennes
- James Cussens, York
- Pascal Denis, INRIA Lille
- Christophe Gonzalès, LIP-6 Paris
- Hachem Kadri, University of marseille
- André Martins, University of Priberam
- Alessandro Moschitti, University of Trento and Qatar Computing Research Institute
- Jian-Yun Nie, University of Montréal
- Ariadna Quattoni, Xerox
- Christrine Sinoquet, University of Nantes
- Andreas Vlachos, University of Sheffield

Local Information

The workshop takes place in **Amphitheater Schwartz**, building 1R3, Université Paul Sabatier (campus de Rangueil). The map of the campus is available at: `http://www.ups-tlse.fr/html/carte.pdf` The course will also take place in **Amphitheater Schwartz**.

Contact

`stergos.afantenos@irit.fr`
`mathieu.serrurier@irit.fr`
`loubes@math.univ-toulouse.fr`



Schedule

Wednesday, December 9th		
9:00	**André Martins**	Parsing as Reduction
10:00		**coffee break**
10:15	**Alessandro Moschitti**	Efficient Structural Kernels for Natural Language Processing
11:15	**Andreas Vlachos**	Imitation learning for structured prediction in NLP
12:15		**Lunch break**
14:00	**Christophe Gonzalès**	Learning Non-stationary Dynamic Bayesian Networks
15:00		**coffee break**
15:15	**Hachem Kadri**	Kernel Methods for Structured Outputs
16:15	**Eustasio del Barrio**	A contamination model for approximate stochastic order
17:15		End
Thursday, December 10th		
9:00	**Xavier Carreras**	Low-rank Matrix Learning for Compositional Objects, Strings and Trees
10:00		**coffee break**
10:15	**Pascal Denis**	Learning Latent Trees for Joint Coreference Resolution and Anaphoricity Detection
11:15	**Ariadna Quattoni**	Hankel Based Methods for Learning Non-Deterministic Automata
12:15		**Lunch break**
14:00	**James Cussens**	Integer programming for Bayesian network structure learning
15:00		**coffee break**
15:15	**Christine Sinoquet**	Modeling of high-dimensional and spatially correlated data with forest of latent tree models
16:15	**Francois Coste**	Learning the language of protein sequences
17:15	**Jian-Yun Nie**	
18:15		End
Friday, December 11th		
9:00	**Alexandre Allauzen**	Deep learning part 1
10:30		**coffee break**
10:45	**Alexandre Allauzen**	Deep learning part 2
12:15		**Lunch break**
14:00	**Alexandre Allauzen**	Hands-on computer session: Deep learning
16:00		End

Machine learning

Centre International de Mathématiques et d'Informatique
TOULOUSE

Alexandre Allauzen

Alexandre Allauzen is assistant Professor at the Université Paris-Sud XI, He is a reseacher at LIMSI-CNRS in the Spoken Language Processing group. His main research interests are : Statistical Machine Translation, Statistical language modeling, Machine learning and natural language processing, Automatic Speech Recognition.

Tutorial : Deep learning and natural language processing

The goal of deep learning is to explore how computers can take advantage of data to develop features and representations appropriate for complex inference tasks. This tutorial aims to cover the basic motivation, ideas, models and learning algorithms in deep learning as well as their application to Natural Language Processing.

First, the tutorial presents the basics of neural networks and the training algorithms via backpropagation. Then, different applications such as POS tagging, language modeling and machine translation will be discussed to introduce more advanced neural architectures. This tutorial also includes historical remarks on the novelty of the deep learning approach and proposes some research perspectives.

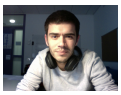
Eustasio del Barrio

Eustasio del Barrio is professor of statistic at the university of Valladolid.

A contamination model for approximate stochastic order

Stochastic ordering among distributions has been considered in a variety of scenarios. Economic studies often involve research about the ordering of investment strategies or social welfare. However, as noted in the literature, stochastic orderings are often a too strong assumption which is not supported by the data even in cases in which the researcher tends to believe that a certain variable is somehow smaller than other. Instead of considering this rigid model of stochastic order we propose to look at a more flexible version in which two distributions are said to satisfy an approximate stochastic order relation if they are slightly contaminated versions of distributions which do satisfy the stochastic ordering. The minimal level of contamination that makes this approximate model hold can be used as a measure of the deviation of the original distributions from the exact stochastic order model. Our approach is based on the use of trimmings of probability measures. We discuss the connection between them and the approximate stochastic order model and provide theoretical support for its use in data analysis, including asymptotic distributional theory as well as non-asymptotic bounds for the error probabilities of our tests. We also provide simulation results and a case study for illustration.

Xavier Carreras



Xavier Carreras is a research scientist at Xerox Research Centre Europe XRCE. His research is in Natural Language Processing and Machine Learning. He is interested in grammatical induction and parsing methods for syntactic-semantic analysis and translation of natural languages.

Low-rank Matrix Learning for Compositional Objects, Strings and Trees

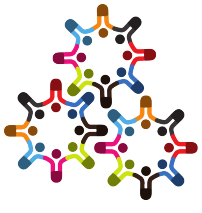
Compositional structures abound in NLP, ranging from bilexical relations, entities in context, sequences and trees. The focus of this talk is in latent-variable compositional models for structured objects that: (1) induce a latent n-dimensional representation for each element of the structure; and (2) learn operators for composing such elements into structures. I will present a framework based on formulating the learning problem as low-rank matrix learning, where we employ the well-known nuclear norm regularizer to favor parameter matrices that are low-rank. I will then illustrate applications of this framework in NLP tasks. First I will show a method to learn word embeddings tailored for specific linguistic relations, which results in word vectors that are both very compact and predictive. Then I will show the importance of low-rank regularization in conjunctive feature spaces, and specifically how our method can propagate weights to conjunctions not observed during training, which results in large improvements in a named entity extraction task. Finally I move to structure prediction and show that low-rank matrix learning can be used to induce the states of weighted automata for sequence tagging tasks.

François Coste

François Coste is research scientist in the dyliss team (Irisa / Inria Rennes). His research is about algorithms learning automatically the syntax of (a set of related) biological sequences.

Learning the language of protein sequences

The linguistic metaphor has been used for long in Molecular Biology in the study of biological sequences. In this talk, I will focus on modelling protein sequences by formal grammars, presenting classical methods from the Bioinformatics field and our recent work to learn automatically the topology of these grammars.



James Cussens

James Cussens is senior lecturer at university of York since 1997. Before tha,t he was researcher at King's College, London, Glasgow Caledonian and Oxford. His research interrest are machine learning particular Bayesian methods and graphical models, applying discrete optimisation algorithms to machine learning, statistical methods in artificial intelligence.

Integer programming for Bayesian network structure learning

With complete data and appropriately chosen parameter priors the problem of finding a Bayesian network with maximal log marginal likelihood (LML) becomes a purely discrete problem: search for a directed acyclic graph (DAG) with maximal LML. We solve this problem of discrete optimisation using integer linear programming (ILP). In many cases this allows us to solve the problem: we find a DAG which we know to have maximal LML. Also using ILP allows prior knowledge, such as known conditional independence relations, to be expressed as constraints on DAG structure. This approach has proven to be particularly successful when learning pedigrees from genetic marker data where it is possible to learn guaranteed maximal LML DAGs with over 1000 nodes. However, many challenges remain, for example when there is missing data and when there is linkage between markers.

Pascal Denis

Pascal Denis is a tenured Research Scientist with INRIA, the French National Institute for Research in Computer Science and Control. His current affiliation is with the Magnet team, a joint research group between INRIA Nord Lille Europe Research Center and University of Lille Computer Science Department (LIFL). The main research focus of this team is on Machine Learning for Information Networks. His main specialization is in statistical Natural Language Processing (NLP), with an emphasis on modeling semantic and discourse-based problems (coreference resolution, temporal and rhetorical structure prediction). Over the recent years, he has become very interested in Machine Learning, and especially in applying ML techniques to NLP problems, as well as to graph structured data.

Learning Latent Trees for Joint Coreference Resolution and Anaphoricity Detection

Noun Phrase Coreference Resolution is one of the most challenging tasks in Natural Language Processing with numerous applications within and outside NLP. After briefly reviewing limitations of previous approaches, this talk introduces a new structured model for learning coreference resolution and anaphoricity detection in a joint fashion. Specifically, we use a latent tree to represent the full coreference and anaphoric structure of a document at a global level, and we jointly learn the parameters of the two models using a version of the structured perceptron algorithm. Our joint structured model is further refined by the use of pairwise constraints which help the model to capture accurately certain patterns of coreference. Our experiments on the CoNLL-2012 dataset, the main benchmark for this task, show large improvements compared to various competing architectures.

Christophe Gonzalès

Christophe Gonzalès is full professor at Paris 6 university. He is currently conducting my research activities in the Decision Theory team, "DEcision making, Intelligent Systems and Operations Research" department of the Computer Science research lab of the university, a.k.a. LIP6. He still works from time to time on additive utility functions, but currently his main research field concerns graphical models for Decision Making, and especially Bayesian and GAI networks.

Learning Non-stationary Dynamic Bayesian Networks

Dynamic Bayesian networks (DBN) are a popular framework for managing uncertainty in time-evolving systems. Their efficient learning has thus received many contributions in the literature. But, most often, those assume that the data to be modeled by a DBN are generated by a stationary process, i.e., neither the structure nor the parameters of the BNs evolve over time. Unfortunately, there exist real-world problems where such a hypothesis is highly unrealistic, e.g., in video event recognition, social networks or road traffic analysis. In this talk, we will review the state-of-the art algorithms for learning "non-stationary DBNs" and propose a new principled approach to learn both the structure and parameters of "non-stationary DBNs".

Hachem Kadri

Since September 2012, Hachem Kadri is an Assistant Professor in the Department of Computer Science at Aix-Marseille University. He is also a member of the Qarma (Machine Learning and Multimedia) team which is situated within the Fundamental Computer Science laboratory (Laboratoire d'Informatique Fondamentale de Marseille, a.k.a. LIF). His research interests lie principally in the fields of machine learning, statistics, and signal processing, more precisely in: kernel methods, functional data analysis, and speech and audio processing.

Kernel Methods for Structured Outputs

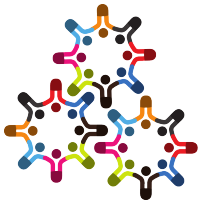
Kernel methods are a popular family of statistical learning methods that exploit training data through the implicit definition of a similarity between data points. This talk will start by explaining the principles of this class of learning algorithms and then discussing some of their advantages and disadvantages. It will then explain how these methods can be extended to handle structured outputs. The talk will concentrate on conceptual rather than technical issues.

André Martins

André Martins is a Research Scientist at Priberam Labs in Lisbon, Portugal. He works on natural language processing and machine learning. He is also affiliated with the Instituto de Telecomunicaçes, at Instituto Superior Técnico.

Parsing as Reduction

In this talk, I will show how we can reduce phrase-based parsing to dependency parsing. Our reduction is grounded on a new intermediate representation, "head-ordered dependency trees," shown to be isomorphic to constituent trees. By encoding order information in the dependency labels, we show that any off-the shelf, trainable dependency parser can be used to produce constituents. When this parser is non-projective, we can perform discontinuous parsing in a very natural manner. Despite the simplicity of this



approach, experiments show that the resulting parsers are on par with strong baselines, such as the Berkeley parser for English and the best non-reranking system in the SPMRL-2014 shared task. Results are particularly striking for discontinuous parsing of German, where we surpass the current state of the art by a wide margin. If time permits, I will describe TurboParser, a multilingual dependency parser based on linear programming.

Alessandro Moshitti

Alessandro Moschitti is a Principal Research Scientist of the Qatar Computing Research Institute (QCRI), within the Hamad Bin Khalifa University and a professor at the CS Department of the University of Trento, Italy. He obtained his PhD in NLP from the University of Rome in 2003. He has worked as an associate researcher at the University of Texas at Dallas, as a visiting professor at the Universities of Columbia, Colorado and John Hopkins and as a visiting research scientist at the IBM Watson Research center of NY and MIT-CSAIL.
His expertise concerns theoretical and applied machine learning in the areas of NLP, Information Retrieval and Data Mining. He has devised innovative kernels within support vector and other kernel-based machines for advanced syntactic/semantic processing, documented in about 220 scientific articles published in major venues. He has been the General Chair of EMNLP 2014 and the Program co-Chair of CoNLL 2015. He has been PI for several European and USA projects. He is currently the PI (QCRI side) of a collaboration project between MIT-CSAIL and QCRI. He has received four IBM Faculty awards, one Google Faculty award, five best paper awards and a best researcher award from the Trento University.

Efficient Structural Kernels for Natural Language Processing
Almost all Natural Language Processing (NLP) applications deal with syntactic and semantic structures, requiring therefore the use of models for processing structured input and generating structured output. Given the complexity of natural language, heuristic methods demonstrate often to be inadequate for encoding such structures in NLP systems, also requiring a considerable design effort. Statistical machine learning approaches for structured output prediction are an effective alternative to the approach above. However, to our knowledge, there are no efficient global inference algorithms enabling the use of structures in the input, in the form of structural kernels, and outputting structures. An efficient approach consists in two steps: (i) the generation of a list of hypotheses of the structured output and (ii) the application of learning to rank approaches based on kernels for selecting the best hypothesis of the list. This talk will introduce the most advanced structural kernels for NLP and then will show methods for efficient linguistic structure prediction for several applications, ranging from concept segmentation and labeling to predicate argument and discourse structures.

Jian-Yun Nie

Jian-Yun Nie is full computer science at university of Montréal His research focuses on information retrieval and on Web search engines. The goal is the improvement of the state of the art and the current practices in this field, through the development of novel information retrieval models, and by exploiting new data sources. These sources, such as user logs, Wikipedia entries and thesauri are put to use to expand, rewrite and otherwise reorganize user queries. His research interests also lie in taking into account the user's various intentions in different application contexts.

How can NLP help IR?
It is often believed that Natural Language Processing (NLP) plays an important role in Information Retrieval (IR). However, the current state of the art of IR does not use extensively NLP techniques. IR has developed its own methods to deal with languages. In this talk, I will review some of the successful and unsuccessful attempts on utilizing NLP in IR. It will be shown that successful uses generally require adaptation of NLP techniques to IR tasks. Two such adaptations will be presented in more detail: query parsing and deep learning in IR. In both cases, the adapted methods improved IR.

Ariadna Quattoni

Ariadna Quattoni is a research scientist at Xerox Research Centre Europe XRCE.

Hankel Based Methods for Learning Non-Deterministic Automata
There is an increasing interest in spectral methods for learning latent-variable language models in the form of weighted automata and context-free grammars. Spectral methods provide an algebraic formulation of the learning problem that directly exploits the recurrence relations satisfied by these functions. I will review the spectral method from an algebraic perspective that exploits Hankel matrices to represent functions. The key idea is that if a function is computed by a finite state process then its corresponding Hankel matrix will be of low rank. Using this fact we can reduce the problem of searching over the space of finite state functions to searching over the space of low-rank Hankel matrices.

Christine Sinoquet

Christine Sinoquet is an Assistant Professor in the Department of Computer Science at Nantes university and researcher at the LINA. Her research interests are : Modeling of high-dimensional and spatially correlated data, Advanced models in machine learning for advanced GWAS strategies, Simulation of realistic genetical data, Recovery of epistatic patterns of susceptibility in complex diseases.

Modeling of high-dimensional and spatially correlated data with forest of latent tree models: application to the modeling of genetical data for genome-scale multilocus association studies.

Machine learning

Centre International
de Mathématiques et d'Informatique
TOULOUSE

At the corner of graph and probability theory, we have proposed a new class of probabilistic graphical model. This model, named FLTM (Forest of latent tree models), is dedicated to the modeling of complex high-dimensional and spatially correlated data. This modeling is made easier by the conception of a scalable learning algorithm. This breakthrough has made possible the modeling of genetic data that describes a population at the genomic level. In a first time, we will describe our learning algorithm. Then, we focus on the estimation of the impact of the choice of the clustering method associated with this algorithm.

Andreas Vlachos

Andreas Vlachos is a lecturer at the University of Sheffield, working on the intersection of Natural Language Processing and Machine Learning. His current projects include natural language generation, automated fact-checking and imitation learning. He has also worked on semantic parsing, language modelling, information extraction, active learning, clustering and biomedical text mining.

Imitation learning for structured prediction in NLP

Imitation learning is a learning paradigm originally developed to learn robotic controllers from demonstrations by humans, e.g. autonomous helicopters from pilot's demonstrations. Recently, algorithms for structured prediction were proposed under this paradigm and have been applied successfully to a number of tasks such as information extraction, coreference resolution, semantic parsing and dynamic feature selection. Key advantages are the ability to handle large output search spaces and to learn with non-decomposable loss functions. In this talk I will discuss the main ideas behind imitation leaning, and describe in detail its application to semantic parsing.

About the CIMI Machine Learning thematic trimester

From September to December 2015, the International Centre for Mathematics and Computer Science (CIMI) organizes a thematic trimester on Machine Learning.

The goal of this trimester is to propose a series of scientific and pedagogical events reflecting common interests of the two laboratories that founded CIMI: Institut de Mathématiques de Toulouse (IMT) and Institut de Recherche en Informatique de Toulouse (IRIT).

The trimester started with a summer school and a first workshop; three other events are due until December:

- Summer school - September 14 to 18, 2015
- Workshop 1 - Optimization in machine learning, vision and image processing - October 6 and 7, 2015
- Workshop 2 - Sequential Learning and Applications - November 9 and 10, 2015
- Big Data Days, November 16 and 17, 2015
- **Workshop 3 - Learning with Structured Data and Natural Language - December 9 to 11, 2015**

Scientific committee

Francis Bach, École Normale Supérieure
Sébastien Bubeck, Microsoft Research
Nicolo Cesa-Bianchi, Università degli Studi di Milano
Rémi Gribonval, lRISA Rennes
Marc Sebban, University Jean Monnet
Noah Smith, Carnegie Mellon University
Johan Suykens, KU Leuven
Marc Teboulle, Tel-Aviv University

Program committee

Stergos Afantenos
Alain Berro
Nicolas Couellan
Aurélien Garivier
Sébastien Gerchinovitz
Jean-Michel Loubès
François Malgouyres
Josiane Mothe
Sandrine Mouysset
Mathieu Serrurier

About CIMI

CIMI stands for Centre International de Mathématiques et Informatique de Toulouse and it is one of the Excellence projects selected by the ANR for the period 2012-2020. CIMI brings together the teams of the Institut de Mathématiques de Toulouse (IMT) and the Institut de Recherche en Informatique de Toulouse (IRIT).

It aims at becoming an international reference in mathematics, computer science and their interactions. The program will attract high-level scientists and students from around the world. It includes actions towards attractiveness, such as Excellence Chairs for long-term visitors, grants for Doctoral and Post-Doctoral students, as well as fellowships for Master students. Attractiveness is further enhanced with thematic trimesters organized within CIMI on specific topics including courses, seminars and workshops.

The innovative tools developed at CIMI will also have a strong economic impact on the region and will profit its industrial partners in major technological areas, making CIMI a strategic partner of the social and economic world.