2017-00186 - Machine learning models from temporal data (Ph.D. thesis)

Level of qualifications required: Graduate degree or equivalent

Fonction: PhD Position

About Inria
Inria, the French National Institute for computer science and applied mathematics, promotes “scientific excellence for technology transfer and society”. Graduates from the world’s top universities, Inria’s 2,700 employees rise to the challenges of digital sciences. With its open, agile model, Inria is able to explore original approaches with its partners in industry and academia and provide an efficient response to the multidisciplinary and application challenges of the digital transformation. Inria is the source of many innovations that add value and create jobs.

About the research centre or Inria department
Located at the heart of the main national research and higher education cluster, member of the Université Paris Saclay, a major actor in the French Investments for the Future Programme (Idex, LabEx, IRT, Equipex) and partner of the main establishments present on the plateau, the centre is particularly active in three major areas: data and knowledge; safety, security and reliability; modelling, simulation and optimisation (with priority given to energy).

The 450 researchers and engineers from Inria and its partners who work in the research centre’s 31 teams, the 100 research support staff members, the high-level equipment at their disposal (image walls, high-performance computing clusters, sensor networks), and the privileged relationships with prestigious industrial partners, all make Inria Saclay Île-de-France a key research centre in the local landscape and one that is oriented towards Europe and the world.

Context
This Ph.D. candidate position is offered at Inria Saclay IdF (https://www.inria.fr/en/centre/saclay) in the LIFEWARE project-team (http://lifeware.inria.fr). This team works in computational systems biology and develops the Biochemical Abstract Machine (BIOCHAM http://lifeware.inria.fr/biocham4) software for modeling, analyzing and now synthesizing biochemical reaction networks (CRNs) using methods from fundamental computer science, mathematics and recently machine learning. The software developments are expected to be integrated in BIOCHAM.

The thesis will be supervised by François Fages who has supervised 29 Ph.D. theses in his career, the last defended in May 2016, and who currently is 100% available with no Ph.D. student.

Assignment
Machine learning is one of the fastest growing area of computer science with far-reaching applications. In scientific applications such as bioinformatics and medicine, the most successful deep learning algorithms are able to make accurate predictions, but with currently no explanation capabilities. Our ambitious approach to these limitations is to investigate learning methods for learning dynamical models from temporal data, instead of directly predictions.

The PAC learning framework of Leslie Valiant provides a basic theory of learnability...
and a nice probabilistic quantification of the errors done when learning Boolean functions. In a first article we have applied this framework to the learning of molecular regulatory networks from time series data, and shown the space-time tradeoff there is between the variety of initial conditions and the length of the time horizon to distinguish causal from correlational events:


The ultimate goal of the Thesis is to design machine learning algorithms well suited to learning reaction networks (and/or influence networks) from noisy data time series coming from real biological experiments, possibly in an active learning framework for proposing realistic experiments.

Further reading:

Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: from Theory to Algorithms. Cambridge University Press. 2014.

Main activities

Development of the theory of learnability and probabilistic estimation of the errors in this context will be an essential part of the work.

The software developments will be expected to be integrated in BIOCHAM, and evaluated on our own models and on the large set of models from the BioModels repository, with time series data produced by stochastic simulations, before considering real data.

Interactions with biologists doing experiments concerning active learning are expected in the course of the Thesis.

Skills

This subject requires common and basic knowledge on probabilities, Boolean logic, graph theory and complexity. The candidate will have to be fluent in English.

There is no specific prerequisite for this internship. However, some specific knowledge on either machine learning, dynamical systems, Prolog, systems biology, synthetic biology, or chemistry will be a plus.

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours
- Sports facilities

Remuneration

1st and 2nd year : 1.982 euros/month (gross salary)

3rd year : 2.085 euros/month (gross salary)