IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Paris

Activity Report 2014

Project-Team CLASSIC

Computational Learning, Aggregation, Supervised Statistical, Inference, and Classification

IN COLLABORATION WITH: Département de Mathématiques et Applications (DMA)

RESEARCH CENTER
Paris - Rocquencourt

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team CLASSIC

Keywords: Machine Learning, Statistical Learning, Sequential Learning, Game Theory, Classification, Information Theory

Creation of the Team: 2009 July 01*, updated into Project-Team:* 2010 January 01, end of the Project-Team: 2014 December 31.

1. Members

Research Scientist
Olivier Catoni [Team leader, CNRS, Senior Researcher, HdR]

Faculty Member
Gérard Biau [Univ. Paris VI, Professor, HdR]

Administrative Assistants
Marine Meyer [CNRS]
Hélène Milome [Inria]
Lindsay Polienor [Inria]

Others
Ilaria Giulini [ENS Paris]
Thomas Mainguy [ENS Paris, until Aug 2014]
Vincent Rivoirard [Univ. Paris IX, Professor, HdR]

2. Overall Objectives

2.1. Overall Objectives

During its last year, the team was reduced. Olivier Catoni and his two PhD students focussed on the study of new statistical models for corpus linguistics and on dimension free bounds for the estimation of the Gram matrix of an i.i.d. sample (possibly in an infinite dimensional Hilbert space) and its application to Principal Component Analysis.

We recall hereafter the themes that were more broadly studied during the lifespan of the project.

We are a research team on machine learning, with an emphasis on statistical methods. Processing huge amounts of complex data has created a need for statistical methods which could remain valid under very weak hypotheses, in very high dimensional spaces. Our aim is to contribute to a robust, adaptive, computationally efficient and desirably non-asymptotic theory of statistics which could be profitable to learning.

Our theoretical studies bear on the following mathematical tools:

- regression models used for supervised learning, from different perspectives: the PAC-Bayesian approach to generalization bounds; robust estimators; model selection and model aggregation;
- sparse models of prediction and ℓ_1–regularization;
- interactions between unsupervised learning, information theory and adaptive data representation;
- individual sequence theory;
- multi-armed bandit problems (possibly indexed by a continuous set);
- statistical modeling applied to linguistics, statistical inference of grammar models.

We are involved in the following applications:

- the improvement of prediction through the on-line aggregation of predictors, with an emphasis on the forecasting of air quality, electricity consumption, production data of oil reservoirs, exchange rates;
- natural image analysis, and more precisely the use of unsupervised learning in data representation;
- computational linguistics;
- statistical inference on biological and neurobiological data.

3. Research Program

3.1. Regression models of supervised learning

The most obvious contribution of statistics to machine learning is to consider the supervised learning scenario as a special case of regression estimation: given n independent pairs of observations (X_i, Y_i), $i = 1, \cdots, n$, the aim is to "learn" the dependence of Y_i on X_i. Thus, classical results about statistical regression estimation apply, with the caveat that the hypotheses we can reasonably assume about the distribution of the pairs (X_i, Y_i) are much weaker than what is usually considered in statistical studies. The aim here is to assume very little, maybe only independence of the observed sequence of input-output pairs, and to validate model and variable selection schemes. These schemes should produce the best possible approximation of the joint distribution of (X_i, Y_i) within some restricted family of models. Their performance is evaluated according to some measure of discrepancy between distributions, a standard choice being to use the Kullback-Leibler divergence.

3.1.1. PAC-Bayes inequalities

One of the specialties of the team in this direction is to use PAC-Bayes inequalities to combine thresholded exponential moment inequalities. The name of this theory comes from its founder, David McAllester, and may be misleading. Indeed, its cornerstone is rather made of non-asymptotic entropy inequalities, and a perturbative approach to parameter estimation. The team has made major contributions to the theory, first focussed on classification [6], then on regression [1] and on principal component analysis of a random sample of points in high dimension. It has introduced the idea of combining the PAC-Bayesian approach with the use of thresholded exponential moments [7], in order to derive bounds under very weak assumptions on the noise.

4. New Results

4.1. Corpus linguistics and Markov substitute processes

Thomas Mainguy and Olivier Catoni studied a new statistical model for natural language modeling, called Markov substitute processes. This model is based on a set of conditional independence properties that are more general than the Markov field assumption. It has connections with context free grammars and forms a collection of exponential families having for this reason nice estimation properties.

4.2. Kernel Principal Component Analysis and spectral clustering

Ilaria Giulini and Olivier Catoni continued their study of dimension free bounds for the estimation of the Gram matrix and more generally for the estimation of the expectation of a random symmetric matrix from an i.i.d. sample. This study, using PAC-Bayes bounds, both leads to new robust estimators with applications to Principal Component Analysis in high of even infinite dimension, and new bounds for the usual empirical Gram matrix estimate. Getting dimension free bounds is important to get new results on Kernel PCA. Applications were also studied to density estimation and to spectral clustering.

5. Partnerships and Cooperations

5.1. National Initiatives

ANR project in the blank program: Calibration (2012–2015; involves Vincent Rivoirard, who is the coordinator; see https://sites.google.com/site/anrcalibration/home)

6. Dissemination

6.1. Teaching - Supervision - Juries

6.1.1. Teaching

E-learning

Visio-conferencing at IFCAM (Indo-French Centre for Applied Mathematics), Summer School on Applied Mathematics, Indian Institute of Science, Bangalore (July 2014). Olivier Catoni gave a three hour presentation on PAC-Bayes bounds applied to statistical learning. The conference is still available on the author's web page.

6.1.2. Supervision

- PhD : Thomas Mainguy, Markov Substitute Processes, a statistical model for linguistics, Université Pierre et Marie Curie, supervised by par Olivier Catoni, (defended on December 11, 2014).
- PhD in progress : Ilaria Giulini, data analysis in high dimension, started in September 2012, Olivier Catoni.

7. Bibliography

Major publications by the team in recent years

[1] J.-Y. AUDIBERT, O. CATONI. *Robust linear least squares regression*, in "The Annals of Statistics", 2011, vol. 39, n^{o} 5, pp. 2766-2794, http://hal.inria.fr/hal-00522534

[2] K. BERTIN, E. LE PENNEC, V. RIVOIRARD. *Adaptive Dantzig density estimation*, in "Annales de l'IHP, Probabilités et Statistiques", 2011, vol. 47, n^{o} 1, pp. 43–74, http://hal.inria.fr/hal-00381984/en

[3] G. BIAU, L. DEVROYE, G. LUGOSI. *Consistency of random forests and other averaging classifiers*, in "Journal of Machine Learning Research", 2008, vol. 9, pp. 2015–2033

[4] G. BIAU, L. DEVROYE, G. LUGOSI. *On the performance of clustering in Hilbert spaces*, in "IEEE Transactions on Information Theory", 2008, vol. 54, pp. 781–790

[5] O. CATONI. *Statistical Learning Theory and Stochastic Optimization — Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, Lecture Notes in Mathematics, Springer, 2004, vol. 1851, 269 pages

[6] O. CATONI. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, IMS Lecture Notes Monograph Series, Institute of Mathematical Statistics, 2007, vol. 56, 163 p. , http://dx.doi.org/10.1214/074921707000000391

[7] O. Catoni. *Challenging the empirical mean and empirical variance: A deviation study*, in "Annales de l'Institut Henri Poincaré - Probabilités et Statistiques", 2012, vol. 48, n^{o} 4, pp. 1148-1185

[8] O. Catoni. *PAC-Bayes bounds for supervised classification*, in "Festschrift in Honour of A. Chervonenkis", A. Gammerman, H. Papadopoulos, V. Vovk (editors), Springer, 2014, pp. 1-15

[9] M. Devaine, P. Gaillard, Y. Goude, G. Stoltz. *Forecasting electricity consumption by aggregating specialized experts; a review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions*, in "Machine Learning", 2012

[10] G. Lugosi, S. Mannor, G. Stoltz. *Strategies for prediction under imperfect monitoring*, in "Mathematics of Operations Research", 2008, vol. 33, pp. 513–528

[11] B. Mauricette, V. Mallet, G. Stoltz. *Ozone ensemble forecast with machine learning algorithms*, in "Journal of Geophysical Research", 2009, vol. 114, http://dx.doi.org/10.1029/2008JD009978

[12] V. Rivoirard, G. Stoltz. *Statistique mathématique en action*, second edition, Vuibert, 2012, http://www.dma.ens.fr/statenaction/

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] T. Mainguy. *Markov Substitute Processes a statistical model for linguistics*, PhD thesis, University Paris 6, Dec, 2014

Other Publications

[14] G. Biau, R. Zenine. *Online Asynchronous Distributed Regression*, July 2014, https://hal.archives-ouvertes.fr/hal-01024673