Workshop on "Challenging problems in Statistical Learning"

January 28-29, 2010
Université Paris 1 Panthéon-Sorbonne
Centre Malher, 9 rue Malher, Paris 4ème, France

Organized by
Charles Bouveyron (SAMM, University Paris 1)
& Gilles Celeux (Select, INRIA Saclay)

With the support of

Introduction

The workshop on "Challenging problems in Statistical Learning" will be held in Paris on January 28-29, 2010. This workshop aims to summarize the new and future problems in statistical learning and to give a good idea of what already exists for dealing with these problems. This two day workshop is split into 4 sessions :

- Session 1 : Statistical learning and complex data
- Session 2 : Learning with social networks
- Session 3 : Regularization and model selection
- Session 4 : New and future problems in statistical learning

Organization

This event is co-organized by the laboratory SAMM of University Paris 1 Panthéon-Sorbonne and the team Select of INRIA Saclay Ile de France.

The workshop is also supported by the Société Française de Statistique.

Dates and place

The workshop will take place in Centre Malher, 9 rue Malher, Paris (4ème) on January 28-29, 2010. The nearest subway station is "Saint Paul" on line n°1.

The talks will be given in the room "Georges Dupuis" in the basement. The coffee-breaks will be held in the rooms 106 (Thursday) and 107 (Friday).

Video lectures

The lectures given during the workshop will be recorded (sound and beamer) by service TICE of university Paris 1 and will be available online at the following address :

http ://epi.univ-paris1.fr/samm-statlearn

Acknowledgements

The organizers would like to thank the following persons and organisms who helped in preparing this workshop :

- The university Paris 1 and its "conseil scientifique"
- INRIA
- Service TICE of university Paris 1
- Aurélien Hazan, Etienne Côme et Omar Aboura
- Katia Evrat and Catherine Girard (INRIA)

Thursday January 28, 2010

9h–12h30 : Session « Statistical learning and complex data »

- 9h00 : F. Murtagh, *Ultrametric wavelet regression of multivariate time series : application to Colombian conflict analysis*
- 10h00 : Coffee-break in room #106
- 10h30 : S. Girard, *On the regularization of Sliced Inverse Regression*
- 11h30 : C. Biernacki, *Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins*

14h–17h30 : Session « Learning with social networks »

- 14h00 : H. Chipman, *Mixed-Membership Stochastic Block-Models for Transactional Data*
- 15h00 : Coffee-break in room #106
- 15h30 : N. Villa, *Visualization of graphs by organized clustering : application to social and biological networks*
- 16h30 : C. Gormley, *A Mixture of Experts Latent Position Cluster Model for Social Network Data*

Friday January 29, 2010

9h–12h30 : Session « Regularization and model selection »

- 9h00 : C. Giraud, *Estimator selection with unknown variance*
- 10h00 : Coffee-break in room #107
- 10h30 : G. Tutz, *Regularization Methods for Categorical Predictors*
- 11h30 : J.-M. Marin, *Importance sampling methods for Bayesian discrimination between embedded models*

14h–17h30 : Session « New and future problems in statistical learning »

- 14h00 : P. Buhlmann, *High-dimensional interventions and causality : some results and many unsolved problems*
- 15h00 : Coffee-break in room #107
- 15h30 : S. Robin, *Statistical analysis of bio-molecular data and combinatorial difficulties : two examples*
- 16h30 : S. Arlot, *Data-driven penalties for optimal calibration of learning algorithms*

Ultrametric wavelet regression of multivariate time series : application to Colombian conflict analysis

Fionn Murtagh

Department of Computer Science
University of London
Email : *fionn@cs.rhul.ac.uk*

We first pursue the study of how hierarchy provides a well-adapted tool for the analysis of change. Then, using a time sequence-constrained hierarchical clustering, we develop the practical aspects of a new approach to wavelet regression. This provides a new way to link hierarchical relationships in a multivariate time series data set with external signals. Violence data from the Colombian conflict in the years 1990 to 2004 are used throughout. We conclude with some proposals for further study on the relationship between social violence and market forces, viz. between the Colombian conflict and the US narcotics market.

On the regularization of Sliced Inverse Regression

Stéphane Girard

Equipe Mistis
INRIA Rhône-Alpes & LJK
Email : *stephane.girard@inrialpes.fr*

Sliced Inverse Regression (SIR) is an effective method for dimension reduction in high-dimensional regression problems. The original method, however, requires the inversion of the predictors covariance matrix. In case of collinearity between these predictors or small sample sizes compared to the dimension, the inversion is not possible and a regularization technique has to be used. Our approach is based on an interpretation of SIR axes as solutions of an inverse regression problem. A prior distribution is then introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that some existing SIR regularizations can enter our framework, which permits a global understanding of these methods. Three new priors are proposed, leading to new regularizations of the SIR method, and compared on simulated data. An application to the estimation of Mars surface physical properties from hyperspectral images is provided.

Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins

Christophe Biernacki

Laboratoire Paul Painlevé
Universite Lille 1
Email : *biernack@math.univ-lille1.fr*

Mixture model-based clustering usually assumes that the data arise from a mixture population in order to estimate some hypothetical underlying partition of the dataset. In this work, we are interested in the case where several samples have to be clustered at the same time, that is when the data arise not only from one but possibly from several mixtures. In the multinormal context, we establish a linear stochastic link between the components of the mixtures wich allows to estimate jointly their parameter – estimations are performed here by Maximum of Likelihood – and to classsify simultaneously the diverse samples. We propose several useful models of constraint on this stochastic link, and we give their parameter estimators. The interest of those models is highlighted in a biological context where some birds belonging to several species have to be classified according to their sex. We show firstly that our simultaneous clustering method does improve the partition obtained by clustering independently each sample. We show then that this method is also efficient in order to assess the cluster number when assuming it is ignored. Some additional experiments are finally performed for showing the robustness of our simultaneous clustering method to one of its main assumption relaxing.

Mixed-Membership Stochastic Block-Models for Transactional Data

Hugh Chipman

Department of Mathematics & Statistics
Acadia University, Canada
Email : *hugh.chipman@acadiau.ca*

Transactional network data arise in many fields. Although social network models have been applied to transactional data, these models typically assume binary relations between pairs of nodes. We develop a latent mixed membership model capable of modelling richer forms of transactional data. Estimation and inference are accomplished via a variational EM algorithm. Simulations indicate that the learning algorithm can recover the correct generative model. We further present results on a subset of the Enron email dataset.

This is a joint work with Mahdi Shafiei.

Visualization of graphs by organized clustering : application to social and biological networks

Nathalie Villa-Vialaneix

Université de Perpignan, IUT de Carcassonne
& Institut de Mathématiques de Toulouse, Université de Toulouse
Email : *nathalie.villa@math.univ-toulouse.fr*

A growing number of applicative fields generate data that are pairwise relations between the objects under study instead of attributes associated to every object : social networks (relations between persons), biology (interactions between genes, proteins), www (relations between websites or blogs), marketing (relations between customers and services)... To help understanding and interpreting such data, specific data analysis tools have been extended from the classical multivariate data analysis : visualization, clustering, classification ...

This talk deals with an exploratory methodology : a common way to help understanding a graph is to cluster its vertices into relevant groups and then to represent the (simplified) graph of clusters. As will be explained, these two objectives (clustering and representation) can be somehow contradictory. Two approaches related to self-organizing maps will be presented and compared on real-world data to solve this issue.

This is a joint work with Fabrice Rossi (LTCI, Télécom ParisTech).

A Mixture of Experts Latent Position Cluster Model for Social Network Data

Claire Gormley

School of Mathematical Sciences (Statistics)
University College Dublin
Email : *claire.gormley@ucd.ie*

Social network data represent the interactions between a group of social actors. Interactions between colleagues and friendship networks are typical examples of such data. The latent space model for social network data locates each actor in a network in a latent (social) space and models the probability of an interaction between two actors as a function of their locations. The latent position cluster model extends the latent space model to deal with network data in which clusters of actors exist – actor locations are drawn from a finite mixture model, each component of which represents a cluster of actors. A mixture of experts model builds on the structure of a mixture model by taking account of both observations and associated covariates when modeling a heterogeneous population. Herein, a mixture of experts extension of the latent position cluster model is developed. The mixture of experts framework allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations. Estimates of the model parameters are derived in a Bayesian framework using a Markov Chain Monte Carlo algorithm. The algorithm is generally computationally expensive – surrogate proposal distributions which shadow the target distributions are derived, reducing the computational burden. The methodology is demonstrated through an illustrative example detailing relations between a group of lawyers in the USA.

Estimator selection with unknown variance

Christophe Giraud

Centre de Mathématiques Appliquées
Ecole Polytechnique
Email : *christophe.giraud@polytechnique.edu*

We consider the problem of Gaussian regression (possibly in a high- dimensional setting) when the noise variance is unknown. We propose a procedure which selects within any collection of estimators $\mathbf{F} = \{\hat{f}_\lambda : \lambda \in \Lambda\}$, an estimator $hat f_{\hat{\lambda}}$ that nearly achieves the best bias/variance trade off. This selection procedure can be used as an alternative to Cross Validation to :

– tune the parameters of a family of estimators
– compare different families of estimation procedure
– perform variable selection.

Regularization Methods for Categorical Predictors

Gerhard Tutz

Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
Email : *g.tutz@gmx.net*

The majority of regularization methods in regression analysis has been designed for metric predictors and can not be used for categorical predictors. A rare exception is the group lasso which allows for categorical predictors or factors. We will consider alternative approaches based on penalized likelihood and boosting techniques. Typically the operating model will be a generalized linear model.

We will start with ordered categorical predictors which unfortunately are often treated as metric variables because software is available. It is shown how difference penalties on adjacent dummy coefficients can be used to obtain smooth effect curves that can be estimated also in cases where simple maximum likelihood methods fail. The difference penalty turns out to be highly competitive when compared to methods often seen in practice, namely simple linear regression on the group labels and pure dummy coding.

In a second step L_1-penalty based methods that enforce variable selection and clustering of categories are presented and investigated. It is distinguished between ordered predictors where clustering refers to the fusion of adjacent categories and nominal predictors for which arbitrary categories can be fused. The methods allow to identify which categories do actually differ with respect to the dependent variable. Finally interaction effects are modeled within the framework of varying coefficients models.

For the proposed methods properties of the estimators are investigated. Methods are illustrated and compared in simulation studies and applied to real world data.

Importance sampling methods for Bayesian discrimination between embedded models

Jean-Michel Marin

Institut de Mathématiques et Modélisation de Montpellier
Université Montpellier 2
Email : *Jean-Michel.Marin@univ-montp2.fr*

We survey some approaches on the approximation of Bayes factors used in Bayesian model choice and propose a new one. Our focus here is on methods that are based on importance sampling strategies, rather than variable dimension techniques like reversible jump MCMC, including : crude Monte Carlo, MLE based importance sampling, bridge and harmonic mean sampling, Chib's method based on the exploitation of a functional equality, as well as a revisited Savage-Dickey's approximation. We demonstrate in this survey how all these methods can be efficiently implemented for testing the significance of a predictive variable in a probit model. Finally, we compare their performances on a real dataset.

This is a joint work with Christian P. Robert.

High-dimensional interventions and causality : some results and many unsolved problems

Peter Bühlmann

Seminar für Statistik
ETH Zürich HG G17, CH-8092 Zurich, SWITZERLAND
Email : *buhlmann@stat.math.ethz.ch*

Understanding cause-effect relationships between variables is of interest in many fields of science. To effectively address such questions, we need to look beyond the framework of variable selection or importance from models describing associations only. We will show how graphical modeling and intervention calculus can be used for quantifying intervention and causal effects, particularly for high-dimensional, sparse settings where the number of variables can greatly exceed sample size.

Statistical analysis of bio-molecular data and combinatorial difficulties : two examples

Stéphane Robin

INRA & AgroParisTech
Email : *Stephane.Robin@agroparistech.fr*

Combinatorial issues are often raised by statistical model inference and selection, in particular when dealing with high-dimensional data. In such cases, asymptotic approximations or Monte-Carlo type methods are often used to approximate the quantities of interest. In this talk, we will present two examples dealing with bio-molecular data. In both of them exacts results can be obtained based on specific combinatorics and algorithmics developments.

We will first consider the typical multiple testing issued that is faced when dealing with high-throughput data. In this framework, most multiple testing procedures require a precise estimation of the proportion of true null hypotheses. This estimation problem can be rephrased as an histogram selection problem, which can be solved via leave-p-out (LpO) cross-validation. We will present explicit results that allow us to manage this model selection problem, avoiding the computational burden inherent to LpO.

We will then consider a segmentation problem encountered when looking for chromosomal aberrations based one microarray data. The detection of breakpoints and the estimation of their number is an old statistical problem. As for the precision of their localisation, only asymptotic results are available. We will present a dynamic programming type algorithm that allows us to explore the whole segmentation space. It provides information on the localisation precision. It furthermore provides a new model selection criterion for the number of breakpoints.

Data-driven penalties for optimal calibration of learning algorithms

Sylvain Arlot

Équipe Willow - LIENS
ENS Paris
Email : *sylvain.arlot@ens.fr*

Learning algorithms usually depend on one or several parameters that need to be chosen carefully. We tackle in this talk the question of designing penalties for an optimal choice of such regularization parameters in non-parametric regression.

First, we consider the problem of selecting among several linear estimators, which includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. We propose a new penalization procedure which first estimates consistently the variance of the noise, based upon the concept of minimal penalty which was previously introduced in the context of model selection. Then, plugging our variance estimate in Mallows' C_L penalty is proved to lead to an algorithm satisfying an oracle inequality.

Second, when data are heteroscedastic, we can show that dimensionality-based penalties are suboptimal for model selection in least-squares regression. So, the shape of the penalty itself has to be estimated. Resampling is used for building penalties robust to heteroscedasticity, without requiring prior information on the noise-level. For instance, V-fold penalization is shown to improve V-fold cross-validation for a fixed computational cost.