Scientific Committee

Peter Bartlett
(University of California, Berkeley)

Evarist Giné
(University of Connecticut)

Vladimir Koltchinskii
(University of New Mexico, Albuquerque)

Gábor Lugosi
(Universitat Pompeu Fabra)

Shahar Mendelson
(Australian National University)

Vitali Milman
(Tel Aviv University)

Steve Smale
(Toyota Technological Institute at Chicago and University of California, Berkeley)

Acknowledgements. The Mathematical Foundations of Learning Theory is supported by the "Ministerio de Ciencia y Tecnología" (ref. BFM2002-11039-E).

Contents

1 Programme

Time	Friday, 18	Saturday, 19	Monday, 21	Tuesday, 22	Wednesday, 23
09:30-10:20	S. Boucheron *Concentration inequalities*	S. Smale *Online learning, mathematical foundations*	P. Long *Generalization of Boosting Algorithms in Theory and Practice*	V. Milman *Unified behavior of convex bodies in high dimensional spaces*	M.Ledoux *On tail and variance inequalities for the largest eigenvalues of random matrix and random growth models*
10:30-11:00	Coffee break				
11:00-11:50	N. Linial *Margins and other related complexity measures*	S. Mendelson *Geometric aspects of Learning Theory*	F. Barthe *Isoperimetry between exponential and Gaussian*	V. Kurkova *Learning from data as an inverse problem*	G.Schechtman *Embedding subsets of Euclidean space in normed spaces*
12:00-12:50	Jon.A.Wellner *Entropy with bracketing for a class on montone functions and some consequences*	D. Foster *Smooth Functional Programming*	J. Zinn *The Central Limit Theorem in Banach spaces and Empirical Processes*	R. Dudley *On some location and scatter functionals*	M. Opper *Tractable approximate inference in probabilistic models*

LUNCH					
15:00-15:50	N.Cesa-Bianchi *Universal prediction, learning with experts, and repeated games*	P. Massart *Optimal model selection for classification*	S. Hart *Adaptive Heuristics*	L. Devroye *Learning and Density estimation*	R. DeVore *Estimators for Supervised Learning*
16:00-16:15	Coffee break				
16:15-17:05	A. Pajor *Geometry of random polytopes*	D.-X. Zhou *Support vector machine soft margin classifiers*	Y. Freund *Average case optimization, driting games and Brownian Motion*	A. Naor *On the dimension of finite metric spaces*	
17:05-17:55	A. Tsybakov *The impact of dimension on classification*	R. Vershynin *Combinatorics of Donsker classes*	A. Barron *Risk Assessment for Model Mixing*	V. Vovk *On-line transduction*	
17:55-18:45	Poster session				

2 General information

Lecture Room: The conference will take place in the "Sala Pere Coromines" of the Institut d'Estudis Catalans (carrer del Carme, 47. Barcelona). The "Sala Nicolau d'Olwer" will be available to lecturers and participants for informal meetings and as computer room.

Secretariat: The conference is organised by the Centre de Recerca Matemàtica, located in the premises of the Universitat Autònoma de Barcelona, in Bellaterra, some 20 Km from Barcelona. A CRM secretary will be available either outside the conference room or in the Sala Nicolau d'Olwer. Please, contact them freely if you need assistance. You can also contact the Centre de Recerca Matemàtica if you have any question or need assistance at other times. The numbers are: 93 581 10 81 (Mrs. Neus Portet and Mrs. Consol Roca).

Grants: Participants receiving grants will be asked to sign some receipt forms the first day of the conference. Please bring a copy of your passport and those receiving travel grants bring a copy of your travel expenses ticket.

Computer facilities: There will be 3 computers available for the participants in room Nicolau d'Olwer.

Breaks: Coffee and pastries will be served during the morning breaks to all participants at the building's ground floor cloister.

Lunch: Five restaurant tickets will be given to all registered participants together with the documentation package for a value of 8€ a day. Tickets will be ready to use during the days of the conference in any of the restaurants provided by the Sodexho pass company. A list of the nearby restaurants will also be provided. We kindly ask you to return any unused tickets to us on the last day of the conference.

Dinner party: A dinner party for all participants has been organized for Tuesday, June 22 at the Restaurant "El Chipiron" (Maremagnum Port Vell, Barcelona), beginning at 8:00 pm. Every participant interested in attending must register the first day of the conference to one of the organizers in charge. There will be a 5€ fee to register and a 35€ for partners to be paid at the moment of registration. No reimbursements will be made.

Picture: A group picture will be taken on the first day of the conference during the coffee break.

Questionnaire: Following the directions of the CRM Governing Board, we give a questionnaire to all the people participating in activities at the CRM in order to assess their level of satisfaction. The questionnaire is anonymous and not mandatory, but we would greatly appreciate if you could answer the questions and return it to us. We thank you for your cooperation.

3 Abstracts of Main Speakers

Concentration inequalities

Stéphane Boucheron, Université Paris-Sud, Orsay

Traditional *exponential inequalities* (Hoeffding/Bernstein) are non-asymptotic counterparts to limit theorems for sums of independent random variables. *Concentration inequalities* are presented as upper-bounds on tail probabilities for functions of many independent random variables.

Concentration inequalities have proved helpful in statistical learning theory (and in Statistics at large) because they are key ingredients in the derivation of *risk bounds* for empirical risk minimizers in classification and bounded regression frameworks.

The most important concentration inequalities are concerned with *suprema of empirical processes*: the most refined versions of Talagrand's concentration inequality for suprema of empirical processes may be considered as process versions of Bernstein inequality. But concentration inequalities also allow to prove that many of the quantities that have been considered in order to quantify the *complexity of a class of functions*, like the empirical VC-dimension, the empirical VC-entropy and (localized) conditional Rademacher averages are sharply concentrated around their mean. This paves the way to data-dependent estimation of the complexity of function classes, which is of great importance in model selection. This tutorial will outline some applications of concentration inequalities to quantities like empirical VC-entropy, or localized Rademacher complexity.

Just as exponential inequalities prove powerless when dealing with sums of poorly integrable independent random variables, using concentration inequalities when dealing with functions of many random variables with unbounded increments may be quite difficult. Hence, in order to get out of the classfication or bounded regression framework, it is useful to get analogue of *moment inequalities* like *Rosenthal-Pinelis inequalities* for sums of not so integrable independent random variables.

In order to fill this need, the tutorial will sketch the recent generalized entropy method. This *generalized entropy method* provides upper-bound on the qth norm of functions of many independent random variables.

The idea is to relate the q-th norm of some functional of independent random variables, with the qth norm of the square root of the Efron-Stein estimate of variance. The recent inequalities derived from the generalized entropy method interpolate between Efron-Stein inequality and modified logarithmic Sobolev inequalities. It is proved by resorting to functionals called ϕ-entropies that generalize both entropy and variance.

References

S. Boucheron, O. Bousquet, G. Lugosi and P. Massart. *Moment inequalities for functions of independent random variables.* Annals of Probability. To appear.

S. Boucheron, G. Lugosi and P. Massart. *Concentration inequalities using the entropy method.* Ann. Probab. **31** (2003), no. 3, 1583–1614.

S. Boucheron, G. Lugosi, P. Massart. *A sharp concentration inequality with applications.* Random Structures Algorithms **16** (2000), no. 3, 277–292.

Email address: stephane.boucheron@lri.fr

Margins and other related complexity measures

Nati Linial, The Hebrew University of Jerusalem

One of the most exciting phenomena in modern theoretical computer science is the interaction between cryptography and the foundations of computer science. So far, it seems that the field of machine learning is more detached from foundational computer science. I will suggest several ways in which such mutually beneficial coopeartion may evolve. I will also present recent work (with Adi Shreibman, Gideon Schechtman and Shahar Mendelson). In this work we investigate several complexity measures of sign matrices, one of them being the margin.

Email address: nati@cs.huji.ac.il

Entropy with bracketing for a class of montone functions and some consequences

Jon A. Wellner, University of Washington, Seattle

It is known that the bracketing entropy of the class of all distribution functions on the real line with respect to the $L_r(Q)$ norm for any $r \geq 1$ and measure Q on $\mathbb{R}$ is bounded by a constant times ϵ^{-1}. The uniform entropy for L_2-norms is also known to be bounded by a constant times ϵ^{-1} via results for convex hulls due to Ball and Pajor, Carl, and van der Vaart and Wellner.

In this talk I will discuss similar results for the class of distribution functions on $\mathbb{R}^d$. In particular I will sketch a proof of the following claim: if $\mathcal{F}_d$ is the class of all distribution functions on $\mathbb{R}^d$, then

$$\log N_{[\,]}(\epsilon, \mathcal{F}_d, L_r(Q)) \le \frac{K}{\epsilon}$$

where $K = K(r, d)$ depends only on r and d and can be computed explicitly. On the other hand, for uniform entropy

$$\sup_Q \log N(\epsilon, \mathcal{F}_d, L_2(Q)) \le \left(\frac{C}{\epsilon}\right)^{\frac{2d}{d+1}},$$

and hence the results differ for $d \ge 2$.

I will also discuss some statistical consequences of these result involving "interval censored" data.

Email address: jaw@stat.washington.edu

Universal prediction, learning with experts and repeated games

Nicolo Cesa-Bianchi, University of Milan

Tutorial on "Universal prediction, learning with experts and repeated games".

Email address: cesa-bianchi@dsi.unimi.it

Geometry of random polytopes

Alain Pajor, Université de Marne-la-Vallée

Random ± 1 polytopes demonstrate the extremal behavior with respect to many characteristics. We illustrate this by showing that the combinatorial dimension, entropy and other parameters of these polytopes are extremal on the whole scale of their argument.

Email address: alain.pajor@univ-mlv.fr

The impact of dimension on classification

Alexandre Tsybakov, University of Paris 6

Classification problems where the dimension of data is high are quite typical. It can be naturally high as in bioinformatics, and it can be artificially high, because of the use of kernel methods. It is interesting to analyze how classification rules behave in asymptotics when the dimension tends to infinity. The talk will discuss the choice of good classification rules in this situation. One of the issues that will be addressed is whether the plug-in rules are outperformed by penalized empirical rules under the margin (low noise) condition when the dimension is high. Another question is how to choose optimally (depending on the dimension) the number of base classifiers for SVM.

Email address: tsybakov@ccr.jussieu.fr

Online learning, mathematical foundations

Steve Smale, Toyota Technological Institute at Chicago and University of California, Berkeley

Shannon sampling is a special case of the general problem of reconstruction of a function from its values at a discrete set of points. The talk will deal with age-old algorithms for solving this problem and new estimates for their error and efficiency.

Email address: smale@math.berkeley.edu

Geometric aspects of Learning Theory

Shashar Mendelson, Australian National University

We will explore some of the connections between Asymptotic Geometry and Learning Theory that have been discovered in the last few years. In particular, we will focus on questions regarding the geometry of coordinate projections of the given class.

Email address: shahar.mendelson@anu.edu.au

Smooth Functional Programming

Dean Foster (with Lyle Ungar), Wharton, University of Pennsylvania

We extend functional programming to handle functions which *smoothly* modify data structures as they are used. In our *Smoothie* language, functions can destructively modify sets, lists, vectors and matrices in ways which guarantee that in the limit of infinite number of calls to the functions, they will effectively converge to returning a unique value and behaving functionally. This allows us to write functions which use "out of date" intermediate results such as partially sorted lists, partially converged numerical algorithms, and data structures which have not been fully updated, while still guaranteeing convergence to true functional behavior. Smoothie also supports non-deterministic function calls and I/O where input may appear in random order or at random times. We present a formal definition of the *asymptotically precise* functions which form the basis of our language, and give examples of their use for problems such as machine learning and database maintenance.

Email address: dean@foster.net

Optimal model selection for classification

Pascal Massart, Université Paris-Sud

Several strategies have been proposed recently to design model selection procedures for classification which are adaptive with respect to margin type conditions. The purpose of this talk is threefold: proposing risk bounds under margin conditions which can serve as benchmarks, showing that "hold-out" can indeed be viewed as a universal margin adaptive selection procedure, proposing data-driven penalized strategies which are based on new concentration inequalities for the empirical risk that seem to be competitive with "hold out" on simulation studies.

Email address: Pascal.Massart@math.u-psud.fr

Support Vector Machine Soft Margin Classifiers

Ding-Xuan Zhou, Department of Mathematics, City University of Hong Kong

Learning Theory studies learning objects from random samples. The purpose of Classification (binary) algorithms is to divide an input (vector) space into two classes, by means of samples.

In this talk we discuss the q-norm soft margin classifier, a support vector machine classification algorithm. We shall provide a PAC error analysis for these algorithms. The misclassification error is bounded by the generalization error associated with a general class of loss functions. The offset is bounded. The convergence rate is then provided by means of the regularization error and the approximation error for reproducing kernel Hilbert spaces. As a corollary we give almost optimal error bounds for weakly separable probability distributions. Even in the well known strictly separable case, our analysis yields the best error bound concerning the influence of the margin. Polynomial kernels and universal kernels are used to demonstrate the main results.

Email address: mazhou@math.cityu.edu.hk

Combinatorics of Donsker classes

Roman Vershynin (with Mark Rudelson), Roman Vershynin (University of California, Davis)

Arguably, the central problem of the theory of empirical processes is the following: *Describe the classes of functions on which the limit theorems of probability hold uniformly.* The classes for which the Central Limit Theorem holds uniformly are known as *uniform Donsker classes.*

Vapnik and Chervonenkis were first to realize that this problem is intimately connected with the combinatorics of the class F. They quantified the size of a class F by what now is called the *Vapnik-Chervonenkis dimension.* This dimension makes sense only for Boolean classes, i.e. for classes of $\{0,1\}$-valued functions. A natural extension of this notion to general classes is possible (although not unique). Arguably, the most suitable extension is the *combinatorial dimension* $v(F,t)$, called by Talagrand "a quantity of fundamental importance". $v(F,t)$ is defined as the maximal size of a subset I of the domain such that find functions from F oscillating in all possible $\pm t/2$ ways on I around some fixed "level" function. This quantity simply measures how much F oscillates.

Theorem 1. *For every uniformly bounded class F,*

$$\int_0^\infty \sqrt{v(F,t)}\, dt < \infty \;\Rightarrow\; F \textit{ is uniform Donsker} \;\Rightarrow\; v(F,t) = O(t^{-2}).$$

The history of the work related to this result goes back to 70's. In 1978, Dudley proved Theorem 1 for Boolean classes of where it clearly reads as follows: F is a uniform Donsker class if and only if its combinatorial dimension $v(F,1)$ is finite. This is one of the main results on the combinatorics of empirical processes on Boolean classes. Talagrand proved Theorem 1 for convex classes, however up to an parasitic factor of $\log^M(1/t)$ in the integrand, and asked about the optimal value of the absolute constant exponent M. Mendelson and Vershynin reduced this exponent to $M = 1/2$. Finally, Theorem 1 is proved by Rudelson and Vershynin. Thus the optimal exponent is $M = 0$.

Theorem 1 is based on a surprising intergal equivalence between the combinatorial dimension and the Koltchinskii-Pollard entropy $D(F,t)$ of arbitrary class:

$$\int_0^\infty \sqrt{D(F,t)}\, dt \approx \int_0^\infty \sqrt{v(F,t)}\, dt$$

where the equivalence is up to an absolute constant factor. Even a pointwise equivalence holds for all regular classes (e.g. for which $v(F,2t) \leq \frac{1}{2}v(F,t)$ for all t):

$$v(F,2t) \lesssim D(F,t) \lesssim v(F,ct) \quad \text{for all } t$$

where $c > 0$ is an absolute constant. For arbitrary classes, this equivalence holds up to the necessary $\log(1/t)$ factor, as proved in 2002 by Mendelson and Vershynin.

Email address: vershynin@math.ucdavis.edu

Generalization of Boosting Algorithms in Theory and Practice

Phil Long, Columbia University

Boosting is a widely successful approach to classification in which a number of "base classifiers" are incorporated into an aggregate classifier. Often, the prediction of the aggregate classifier is a weighted vote over the predictions of the base classifiers. The base classifiers are chosen one at a time in rounds; in a given round, the training examples are re-weighted to assign more importance to "difficult" examples, and a base classifier is chosen in that round to minimize the weighted training error. The generalization of the aggregate classifier has been seen to improve with the number of rounds, even after it has correctly classified all of the training data. This seemed to run counter to the common view that a learning algorithm should strike a balance between the complexity of a hypothesis and its fit to the data. This phenomenon was explained by the "margin analysis," a theoretical analysis that takes account of the margins by which a voting classifier correctly classifies training examples, i.e. how much more weight votes correctly than votes incorrectly. One proof of a margin bound can be paraphrased as saying that a classifier that often achieves large margins is effectively simple, as its behavior can be approximated by a simple majority vote over a few of its base classifiers. This leads to the question: what if a learning algorithm tried to directly fit data using classifiers that are simple in this sense? An algorithm based on this idea performs well on a number of benchmark datasets, especially big ones. But on some datasets, it performs much worse than boosting. This suggests that further theory regarding generalization of boosting algorithms is needed. I will describe one analysis that shows that boosting algorithms take nearly full advantage of situations in which there are a number of base classifiers that are individually fairly accurate, but complement one another, in the sense that they tend to make errors in different places. I will close with a critique of this analysis, and some open questions. (This talk includes joint research with Vinsensius Vega and Sanjoy Dasgupta.)

Email address: plong@cs.columbia.edu

Isoperimetry between exponential and Gaussian

Franck Barthe (joint with P.Cattiaux and C.Roberto),
Université Paul-Sabatier, Toulouse

We are interested in dimension free isoperimetric inequalities for product measures of exponential type. Our results complete the classical Gaussian results as well as the more recent ones by Talagrand and by Bobkov and Houdré. Our approach relies on improving properties of the corresponding semigroups, Sobolev type inequalities and capacity-measure inequalities.

Email address: barthe@math.ups-tlse.fr

The Central Limit Theorem in Banach spaces and Empirical Processes

Joel Zinn, Texas A&M University

On the topic of the Central Limit Theorem in Banach spaces and for Empirical Processes, we'll give a brief description of the goals, previous results and their limitations. We'll then present some new results and the possibilities of extending these results to resolve some of the open problems in this area.

Email address: jzinn@math.tamu.edu

Adaptive Heuristics

Sergiu Hart, Center for Rationality, Department of Mathematics, and Department of Economics The Hebrew University of Jerusalem

We will exhibit a large class of simple rules of behavior, which we call "adaptive heuristics", and show that they generate in the long run rational behavior. These heuristics are based on natural "regret" measures. We will also discuss various multi-person dynamical models – learning, evolutionary, and heuristics – and a basic mathematical tool in this area – Blackwell's Approachability Theory and its extensions. (Joint work with Andreu Mas-Colell, UPF, Barcelona) http://www.ma.huji.ac.il/ hart/abs/adaptdyn.html

Email address: hart@huji.ac.il

Average case optimization, drifting games and Brownian Motion

Yoav Freund, Center for Computational Learning Systems, Columbia University

The computational task that lies in the core of many machine learning problems is the minimization of a cost function called the training error. This problem is frequently solved by local search algorithms such as gradient descent. On the other hand, the cost function can usually be expressed as a sum over many terms, each corresponding to the loss of the model on a single training example. We show that the iteratively minimizing a cost function of this form by local search is closely related to the following game:

Imagine you are a shepherd in charge of a large herd of sheep and your goal is to concentrate the sheep in a particular small area by nightfall. Your influence on the sheep movements is represented by vectors which define the direction in which you want each sheep to move. The lengths of the vectors correspond to the fraction of your "energy" that you spend on moving the particular sheep, and these lengths sum to one. The sheep then have to respond by moving in a way that has a slight correlation with the influence direction on average.

We characterize the min/max solution to this game and show that by taking the appropriate small-step/continuous-time limit, this solution can be characterized by a stochastic differential equation.

By solving this differential equation we re-derive some known boosting and on-line learning algorithms as well as design some new ones with desirable properties.

Email address: freund@cs.columbia.edu

Risk Assessment for Model Mixing

Andrew Barron (with Gilbert Leung), Yale University

We exhibit advantageous risk properties of model mixing compared with model selection. Though some of the principles apply more generally, here we focus on regression with normally distributed errors. The aim is to estimate the vector of means of the response, with sum of squared error loss, and, moreover, when there are several models available, we aim to have a combined estimator for which the risk is not much more than the minimum of the risks of the individual models. Suppose we have an estimator for each model, and following Charles Stein, we have an unbiased estimator of the risk for each model. We mix the estimators across the models using data-dependent weights and evaluate an unbiased estimator of the risk of the mixture. An important case occurs when the estimators are Bayes for some prior for each model and when the mixing weights are posterior probabilities of the models. Then, building on analysis by Ed George, the unbiased estimator of the risk of the mixture is shown to equal the posterior weighted average of the risk estimators for each model plus an additional term for the posterior weighted average of how far the estimates for each model are from the mixture. We show by example that Bayes mixtures, even with as few as two models, can have risk arbitrarily larger than the minimum of the risks of the individual estimators. Alternative mixing weights are given that have more advantageous regret properties. In the case of Bayes estimators for each model, the mixing weights may be taken to be proportional to the square root of a Bayes factor times the prior. Then the unbiased estimator of the risk of the mixture takes the much simpler form of simply the weighted average of the risk estimators, with no additional term. In some settings one can arrange these mixing weights to heavily favor the models with smaller

estimated risks, and then give information-theoretic characterization of the unbiased estimator of the risk of model mixing. This analysis shows that the unbiased estimator of risk of model mixing is not much larger than the minimum of the risk estimates among the models, uniformly over all possible data outcomes. Correspondingly, it is shown that the risk itself (not only its unbiased estimator) inherits this property and is never much larger than the minimum of the individual risks. Examples involve standard subset models for regression with the least squares estimators of the coefficients, and subset models with Stein shrinkage factors. Here the regret is bounded by a small constant times the log of the number of models considered. Alternatively, to deal with large numbers of models with varying complexity, one may incorporate a complexity-based prior and give clean bounds on the mean squared error in terms of the trade-off between individual risk and complexity. Relationships with other methods of model adaptation are discussed.

Email address: andrew.barron@yale.edu and gilbert.leung@yale.edu

---

Unified behavior of convex bodies in high dimensional spaces

V.Milman, Tel Aviv University

I will give examples of results which demonstrate unified behavior of all finite dimensional normed spaces (and finite dimensional convex bodies) to a very high degree of "sameness". These examples will include the behavior of an arbitrary convex body under different symmetrization procedures, random projections, Minkowski additions, formulas for an essential diameter and others.

Email address: milman@post.tau.ac.il

Learning from data as an inverse problem

Vera Kurkova, Institute of Computer Science, Academy of Sciences of the Czech Republic

Learning modelled as minimization of empirical error functionals will be reformulated in terms of the theory of inverse problems. It will be shown that properties of operators defining inverse problems describing learning from empirical data (compactness, form of adjoint) allow to apply results from infinite-dimensional extension of Moore-Penrose theory of pseudoinversion. Several representation theorems describing pseudosolutions and regularized solution of the learning task will be derived from this theory and rates of convergence of regularized solutions with decreasing regularization parameters will be derived. Improvements of stability that can be obtained by regularization will be described in terms of condition numbers of Gram matrices. Optimal solutions described by representer theorems will be compared with suboptimal ones achievable by kernel models with bounded complexity. Estimates of rates of convergence of such suboptimal solutions with increasing model complexity will be derived from theory of nonlinear approximation with variable-basis. The role of smoothness expressed in terms of variation norms will be discussed in connection with the curse of dimensionality.

Email address: vera@cs.cas.cz

On some location and scatter functionals

R. M. Dudley, MIT

Let $\mathcal{P}$ be a set of probability measures on Euclidean space R^d, invariant under each non-singular affine transformation $x \mapsto Ax = Bx + v$ where B is linear and and non-singular and $v \in R^d$. A functional $P \mapsto \mu(P) \in R^d$ on $\mathcal{P}$ is called an affinely equivariant location functional on $\mathcal{P}$ if $\mu(P \circ A^{-1}) = A\mu(P)$ for all non-singular affine A and all $P \in \mathcal{P}$. The classical example is the mean on the class $\mathcal{P}$ of laws having finite mean. For $d = 1$, the median gives an example defined on all laws. A functional $P \mapsto \Sigma(P)$ whose values are nonnegative definite symmetric $d \times d$ matrices is called an affinely equivariant scatter functional on $\mathcal{P}$ if $\Sigma(P \circ A^{-1}) = B\Sigma(P)B'$ for all non-singular affine $A: \ x \mapsto Bx + v$ and all $P \in \mathcal{P}$, as for the covariance matrix of laws with finite second moments. The object is to find such functionals (μ, Σ) defined on as many laws as possible, preferably all laws, and continuous with respect to weak convergence on $\mathcal{P}$. For $d = 1$, one can smoothly trim a law with distribution function F, multiplying it by $g(F(x))$, where g is smooth, $0 \leq g(y) \equiv g(1-y)$, and for some $\delta \in (0, 1/4)$, $g(y) = 0$ for $0 \leq y \leq \delta$ and $g(y) = 1$ for $2\delta \leq y \leq 1 - 2\delta$. Then one can normalize and take the mean and variance of the resulting distribution. For general $d \geq 1$, functionals based on M-estimation for t distributions are applicable. But for $d \geq 2$, an affinely equivariant scatter functional, defined and weakly continuous at all laws, must be identically 0.

Email address: rmd@math.mit.edu

Learning and density estimation

Luc Devroye, McGill University

We show how statistical learning principles aid in the design of density estimates with universal performance guarantees.

Email address: luc@cs.mcgill.ca

How well does a Metric Space Embed into a Normed Space?

Assaf Naor, Microsoft Research

The problem of estimating the least distortion required to embed a finite metric space into a normed space arose in the 1960s in the local theory of Banach spaces. It turns out that several fundamental geometric problems in Banach space theory can be characterized in terms of the embeddability of certain classes of finite metrics. Additionally, due to intensive investigations by computer scientists in the past decade, it is now well understood that embedding results into various normed spaces are a powerful tool in the design of approximation algorithms. Bourgain has shown that every n-point metric space embeds into Euclidean space with distortion O(log n), and this bound is best possible without additional geometric information. In this talk we will present refinements of this result and describe related problems and techniques of modern embedding theory.

Email address: anaor@microsoft.com

On-line transduction

Vladimir Vovk, Royal Holloway

One of the central problems of machine learning is that of "hedged prediction": designing prediction algorithms whose predictions incorporate, or are complemented with, a valid measure of their own reliability. The existing results in this area can be conveniently divided into three big classes, corresponding to the inductive, Bayesian, and transductive styles of prediction. Only recently it has been shown that the errors made by a wide class of transductive prediction algorithms at different trials are independent when these algorithms are used in the on-line mode. This gives a frequentist justification for many old and recent transductive prediction algorithms. In this talk I will review the history of the transductive approach, contrasting it with the two rival approaches, and state some new results.

Email address: vovk@cs.rhul.ac.uk

On tail and variance inequalities for the largest eigenvalues of random matrix and random growth models

Michel Ledoux, University of Toulouse, France

We survey recent results on the asymptotic behavior of the largest eigenvalues of random matrix and random growth models. We ask for the corresponding non-asymptotic tail and variance inequalities. In particular, we consider the possible usefulness of functional tools (hypercontractivity, logarithmic Sobolev inequalities...) to reach sharp bounds at the rate given by the Tracy-Widom asymptotics.

Email address: ledoux@math.ups-tlse.fr

Embedding subsets of Euclidean space in normed spaces

Gideon Schechtman, Weizmann Institute of Science

This is a joint generalization of (Milman's version of) Dvoretzky's theorem and (a recent generalization by Klartag and Mendelson of) the Johnson-Lindenstrauss Lemma. The proof uses a simple application of Talagrand's Majorizing measure theorem.

Email address: gideon@weizmann.ac.il

Tractable Approximate Inference in Probabilistic Models

Manfred Opper, Aston University

Probabilistic models (aka graphical models, Bayes belief networks) provide a powerful framework for modeling statistical dependencies between data

[1]. Their goal is to explain complex observed data by a set of unobserved, hidden random variables based on the joint distribution of both sets of variables. Applications range from medical expert systems over encoders in telecommunication systems to spatial probabilistic models in meteorology. The price that a modeler has to pay for the high degree of flexibility of these models is the vast increase in computational complexity for adapting the models to data and for predicting unobserved causes (hidden nodes) from observations when the number of network nodes is large. Both statistical inference about hidden variables and model training usually requires marginalization of multivariate probability distributions. This leads to the problems of summing up an exponentially large (in the number of random variables) number of terms or of performing high dimensional integrals which usually cannot be done analytically. Since often classical techniques such as Monte Carlo methods can be too time consuming, there has been a lot of effort in developing new approximate inference techniques for Bayes networks [2]. My talk aims at explaining some of the ideas behind such approximations and at demonstrating some of their applications. I will first discuss Variational Approximations in which the true intractable probability distribution is approximated by another one which is optimally chosen within a given, tractable family (often factorizing in all or in groups of random variables). The method produces an exact bound on the marginal probability of the observed data. More recent expectation consistent techniques try to overcome the limitations of factorizing distributions and can be understood from the idea of global consistency between local approximations. Experiments indicate that often the desired marginals are well approximated by such methods. However the computation of theoretical estimates for the accuracy is at present a challenging open problem.

References
[1] M. Jordan, ed.,Learning in Graphical Models, MIT Press (1999).
[2] M. Opper and D. Saad, eds. Advanced Mean Field Methods, Theory and Practice, MIT Press (2001).

Email address: m.opper@aston.ac.uk

Estimators for Supervised Learning

Ronald DeVore, University of South Carolina

The following regression problem occurs in various settings of supervised learning. We are given a domain $X \subset I\!R^d$ and an interval $Y = [-M, M] \subset$

$I\!R$. There is a probability measure ρ defined on $Z := X \times Y$ which is unknown to us. We see information about ρ from samples $\mathcal{Z} := \{(x_i, y_i)\}_{i=1}^m$ which are drawn at random with respect to ρ. We are interested in learning the function $f_\rho(x) := \int\limits_Y y\, d\rho(y|x)$ which is the expected value of y given x.

One way of constructing an estimator for f_ρ is to choose a space of functions $\mathcal{S}$ (called the hypothesis space in learning theory) and find the function $f_{\mathcal{Z}}$ from $\mathcal{S}$ which minimizes the empirical least squares error:

$$f_{\mathcal{Z}} := \underset{f \in \mathcal{S}}{argmin} \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2. \tag{0.1}$$

Cucker and Smale have shown how to estimate the performance of such an estimator by using the Kolmogorov entropy of the set $\mathcal{S}$ in $L_\infty(X)$. They then build estimators in the following way. Given a priori knowledge that f_ρ is in a ball $B(W)$ of a smoothness class W, they choose $\mathcal{S}$ depending on this knowledge and then obtain the estimator $f_{\mathcal{Z}}$. A typical example of their theory gives that if $f_\rho \in B(W)$ with $W = W^r(L_\infty(X))$ then choosing $\mathcal{S}$ as a ball in $W^r(L_\infty(X))$ one obtains

$$E(\|f_\rho - f_{\mathcal{Z}}\|_{L_2(\rho)}) \leq c m^{-\frac{r}{2r+2}}, \quad m = 1, 2, \ldots. \tag{0.2}$$

This talk will center on several ways to improve on these estimators. We show that using tools from approximation theory such as linear or nonlinear widths then we can construct estimators where the right side of (0.2) is replaced by $cm^{-\frac{r}{2r+1}}$ and that this decay rate is optimal. Secondly, the class $W^r(L_\infty(X))$ can be replaced by much larger classes such as the Sobolev classes $W^r(L_p(X))$, $p > d$. Perhaps most importantly we show how to construct estimators that do not require any a priori assumptions on f_ρ but still perform optimaly for a large range of smoothness classes. These latter estimators are built using nonlinear methods of approximation such as adaptive partitioning and n term wavelet approximation.

This work is in collaboration with Peter Binev, Albert Cohen, Wolfgang Dahmen, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov.

Email address: devore@math.sc.edu

4 Abstracts of Posters

Support Vector Machines Capacity Control by Unsupervised Clustering

Davide Anguita, Sandro Ridella, Fabio Rivieccio and Rodolfo Zunino

In the framework of Support Vector Machine (SVM) classifiers [3], an unsupervised analysis of empirical data [1] supports an ordering criterion for the families of possible functions. The proposed approach ultimately refines the basic SRM paradigm by introducing a selection preference within the individual members of a family: an unsupervised analysis of available data supplies the cognitive basis for restricting the family of admissible classifiers. The unsupervised domain analysis follows a two-step process.

Step 1). A feature-space version of the classical k-means algorithm partitions the data sample into two clusters.

Step 2). The pair of data clusters thus obtained support an (artificial) classification problem in which cluster memberships determine pattern classes. This artificial dataset undergoes a conventional SVM training process and the resulting separating hyperplane, which in the following will be denoted by $\mathbf{w}_{KM}$, can be expressed as

$$\mathbf{w}_{KM} = \sum_{i=1}^{np} y'_i \alpha'_i \Phi(\mathbf{x}_i) \qquad (0.1)$$

where the coefficients $\{\alpha'_i\}$ are associated with cluster-related provisional classes, y'_i.

The purpose of the second step is to yield a Support-Vector based representation of the clustering results. These offer a straightforward interpretation: whenever unsupervised clustering matches the actual distribution of pattern classes, the unsupervised separation surface, $\mathbf{w}_{KM}$ and the real classification surface $\mathbf{w}_0$ coincide. Of course, the opposite case may occur in which the target distribution is totally uncorrelated with the obtained clusters.

The proposed method uses the result of unsupervised clustering, $\mathbf{w}_{KM}$, as the central "reference" for the family of admissible classifier configurations. Therefore, the crucial parameter is the distance, $\rho_{\mathbf{w}}$, in the weight space from any weight configuration, $\mathbf{w}$, to the central reference, $\mathbf{w}_{KM}$. The distance-based ordering reduces the number of possible functions in the Maximal-Discrepancy (MD) approach [2], as follows. For a given value of $\rho_{\mathbf{w}}$, only the classifier configurations lying within the sphere of radius $\rho_{\mathbf{w}}$ will be considered in the model selection procedure. The sphere radius, $\rho_{\mathbf{w}}$, is the crucial quantity driving the SRM principle.

In a sample-based classifier design, one is clearly interested in the predicted performance of a classifier trained on the real targets. In the above unsupervised-centered approach, this means that the configuration $\mathbf{w}_0$ delimits the valid portion of the classifier space and all the solutions $\mathbf{w}$ found with the MD estimation process must pass the distance-based selection

$$\rho_{\mathbf{w}} = \|\mathbf{w} - \mathbf{w}_{KM}\| \leq \|\mathbf{w}_0 - \mathbf{w}_{KM}\| \tag{0.2}$$

The effectiveness of the capacity constraint was inspected on two real-world datasets.

The MD estimation procedure involved a series of Montecarlo iterations using random target settings. The weight set, $\mathbf{w}_n$, generated at the n-th iteration was related to the reference $\mathbf{w}_{KM}$ by: $\rho_{\mathbf{w}_n} = \|\mathbf{w}_{KM} - \mathbf{w}_n\|$. The MD value associated with $\mathbf{w}_n$was taken into account in evaluating $\Delta\pi$ whenever the distance $\rho_{\mathbf{w}_n}$satisfied $\rho_{\mathbf{w}_n} \leq \rho_0$. Instead, if $\mathbf{w}_n$ was found to lie outside the bounding sphere, then the function lying closest to $\mathbf{w}_n$ and on the sphere was used to calculate the MD; in those cases, the spherical constraint was fulfilled by straight equality. The adopted MD bound was:

$$\pi = \nu + MD + 3\sqrt{\frac{\ln(1/\delta)}{2 \cdot np}} \tag{0.3}$$

The first testbed was the original Wisconsin Breast Cancer dataset (500 patterns) [4]. The empirical error scored was ν=4.4%. The subsequent Montecarlo estimates involved 100 iterations, and yielded an MD estimate of $\Delta\pi$=3.4% for the unsupervised-constrained SVC. Such a result provided a 10% improvement over the previous best penalty term $\Delta\pi$=4.0%, obtained by the unconstrained SVC according to the general SRM approach. In summary, the overall 95%-confidence bound to the generalization error, resulted in π = 4.4% + 3.4% + 16.4% = 24.2%.

Another experiment on the NIST handwritten-numerals testbed was carried out by randomly picking 500 instances of the "3" and another 500 of the "8" numeral classes. The comparison gave a value $\Delta\pi$=20.9% for the MD estimate without the constraint, which reduced to $\Delta\pi$=3.9% in the spherical-constrained estimation procedure. Thus, the overall 95%-confidence bound yielded a value π = 0.7% + 3.9% + 11.6% = 16.2%. Instead, for the unconstrained MD estimate, the complexity associated with the entire family of classifiers resulted in a generalization bound π=33.2%. This result clearly shows that the presented selection criterion can dramatically improve the performance over that obtained by the classical MD procedure.

References

[1] Lloyd SP "Least squares quantization in PCM", *IEEE Trans. Inf. Theory*, 1982, vol.28, No.2, pp.127-135.

[2] Bartlett P, BoucheronS, Lugosi G "Model selection and error estimation", *Machine Learning*, 2002, vol. 48, No.1-3, pp.85-113.

[3] Vapnik V., *Statistical Learning Theory*, 1998, Wiley-Interscience Pub.

[4] Wolberg WH, Mangasarian OL "Multisurface method of pattern separation for medical diagnosis applied to breast cytology" Proc. Nat. Acad. Sci., U.S.A., December 1990, vol.87, pp 9193-9196.

D.Anguita, S.Ridella, F.Rivieccio and R.Zunino are with the DIBE - Dept. Biophysical and Electronic Engineering - University of Genoa, Via all'Opera Pia 11a - 16145 Genova - Italy.

Email address: anguita@dibe.unige.it, ridella@dibe.unige.it, rivieccio@dibe.unige.it and zunino@dibe.unige.it

---

Statistical Problems of Quantum Homodyne Tomography

L.M.Artiles*, M.I.Guţă†

The fields of Quantum Computation and Quantum Information study the processing of information carried by quantum mechanical devices. In 1994, Shor showed that a practically relevant classical computation problem, namely the factoring of large numbers, with wide applications in communication and cryptography, could be speeded up exponentially on a quantum computer. This has attracted scientists with backgrounds in quantum physics, computer science and information theory [4], and motivated developments in engineering of quantum states and devices able to perform communication and computation tasks, control of its evolution and reading measurements results. It has also motivated the search of areas in which the mathematical machinery of quantum computation might lead to efficient algorithms. There have been already some incursions on this direction in the area of statistical learning [2].

Our purpose is to give an overview of some statistical problems arising from a experimental technique currently applied in optical physics laboratories, namely Quantum Homodyne Tomography [3], which has been recently investigated [1]. In this way we hope to bring Quantum Computation and Quantum Information to the attention of the statistical community, and in particular statistical learning one, as a source of both, interesting problems and powerful techniques.

We describe quantum tomography as an inverse statistical problem. We show results of consistency in different norms, for Pattern Function Projection Estimators as well as for Sieve Maximum Likelihood Estimators for the density matrix of the quantum state and its Wigner function. The density matrix and the Wigner function are two different ways of representing quantum state. Results are derived from concentration inequalities and entropy methods – these statistical techniques are widely used in statistical learning. We illustrate via simulated data the performance of the above mentioned estimators. There remain many open problems, ex. rates of convergence, studying other estimators, etc. Designing appropriate methods which adapt to the data, learn from the data, is our major concern.

References

[1] Artiles, L.M., Gill, R., Guţă, M.I., “An invitation to Quantum Tomography”, *submitted to Journal of the Royal Stat. Soc. B: Methodological,* 2004.

[2] Imre, S., Balázs, F., “Maximum Likelihood Based Quantum Set Separation”, quant-ph/0402081, 2004.

[3] Leonhardt, U., *Measuring the Quantum State of Light*, *Cambridge University Press (1997)*,

[4] Nielsen, M., Chuang, I., *Quantum Computation and Quantum Information*, Cambridge University Press (2000).

*Eurandom, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, †Mathematical Institute, Utrecht University, Box 80010, 3508 TA Utrecht, The Netherlands..

Email address: artiles@eurandom.tue.nl, http://euridice.tue.nl/∼lartiles/
and guta@math.uu.nl, http://www.math.uu.nl/people/guta/

Gibbs classifiers

J. Y. Audibert, Université Paris VI and CREST

This work proposes an efficient way of choosing the temperature of a Gibbs classifier which is based on data-dependent relative error bounds in the PAC-Bayesian framework.

Specifically, we assume that we observe an i.i.d. sample $(X_i, Y_i)_{i=1}^N$ of random variables distributed according to a product probability measure $\mathbb{P}^{\otimes N}$, where $\mathcal{X}$ is the input space, $\mathcal{Y}$ a discrete output space and $\mathbb{P}$ a probability distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F}$ be the model (i.e. a set of prediction functions) and π a probability distribution on $\mathcal{F}$ which is introduced to structure the model. Let $\bar{P} \triangleq \frac{1}{N} \sum_{i=1}^N \delta_{(X_i,Y_i)}$ be the empirical distribution. Let $r(f) \triangleq \bar{P}[Y \neq f(X)]$ and $R(f) \triangleq \mathbb{P}[Y \neq f(X)]$ denote respectively the empirical and expected risks of the prediction function f. Consider the Gibbs estimator which classifies using a function drawn from the distribution $\pi_{-\lambda r}(df) \triangleq \frac{\exp[-\lambda r(f)]}{\mathbb{E}_{\pi(df')} \exp[-\lambda r(f')]} \cdot \pi(df)$, where $\lambda > 0$ is the Gibbs inverse temperature.

This work chooses the parameter λ by successive comparisons. This allows a better variance control. Namely, let $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$. For any $0 \leq j \leq \log N$, introduce $\lambda_j \triangleq 0.19\sqrt{N} e^{\frac{j}{2}}$ and

$$\mathcal{C}(j) \triangleq \log \mathbb{E}_{\pi_{-\lambda_j r}(df)} \exp\Big(\frac{\lambda_j^2}{N} \pi_{-\lambda_j r}(df') \bar{P}[f(X) \neq f'(X)] \Big).$$

Let ϵ be the desired confidence level. For any $0 \leq i < j \leq \log N$, introduce

$$\begin{aligned} S(i,j) \triangleq{}& \frac{\lambda_j}{N} g(\tfrac{\lambda_j}{N})\big(1 + \tfrac{\lambda_j}{2N}\big) \mathbb{E}_{\pi_{-\lambda_i r}(df)} \mathbb{E}_{\pi_{-\lambda_j r}(df')} \bar{P}[f(X) \neq f'(X)] \\ &+ \frac{1}{\lambda_j} \Big[1 + \frac{\lambda_j}{N} g(\tfrac{\lambda_j}{N}) \big(1 + \tfrac{\lambda_j}{2N}\big)^2 \Big] \Big\{ 2\mathcal{C}(i) + 2\mathcal{C}(j) + 3 \log[2 \log^2(eN)\epsilon^{-1}] \Big\}. \end{aligned}$$

Algorithm. *Let $u(0) = 0$. For any $k \geq 1$, define $u(k)$ as the smallest integer $j \in]u(k-1); \log N]$ such that*

$$\mathbb{E}_{\pi_{-\lambda_j r}(df)} r(f) - \mathbb{E}_{\pi_{-\lambda_{u(k-1)} r}(df')} r(f') + S\big[u(k-1), j\big] \leq 0.$$

Classify using a function drawn according to the Gibbs distribution $\pi_{-\lambda_{u(K)} r}$ obtained at the last iteration.

The choice of the temperature is trustworthy since we have: **Theorem.** *Let*

$$\mathbb{G}(j) \triangleq \mathbb{E}_{\pi_{-\lambda_{j-1} R}(df)} R(f) + \frac{\sup_{0 \leq i \leq j} \Big\{ \log \mathbb{E}_{\pi_{-\lambda_i R}(df)} \mathbb{E}_{\pi_{-\lambda_i R}(df')} \exp\Big(\frac{C\lambda_i^2}{N} \mathbb{P}[f(X) \neq f'(X)] \Big) \Big\}}{\lambda_j} + C \frac{\log[\log(eN)\epsilon^{-1}]}{\lambda_j}.$$

With $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, for any $k \in \{1, \ldots, K\}$, we have

- $\mathbb{E}_{\pi_{-\lambda_{u(k)}r}(df)} R(f) \leq \mathbb{E}_{\pi_{-\lambda_{u(k-1)}r}(df)} R(f)$,
- $\mathbb{E}_{\pi_{\lambda_{u(K)}r}(df)} R(f) \leq \min_{1\leq j\leq \log N} \mathbb{G}(j)$.

This guarantee is tight to the extent that, under some variants of Tsybakov's complexity and margin assumptions, it ensures that the previous algorithm achieves the optimal convergence rate in a minimax sense.com

Email address: audibert@ureach.

A PAC-style Model for Learning from Labeled and Unlabeled Data

Maria-Florina Balcan, Avrim Blum

There has recently been substantial interest in practice in using unlabeled data together with labeled data in machine learning, and a number of different approaches have been developed. However, the assumptions these methods are based on are often quite distinct and not captured by standard theoretical models. In this work we describe a PAC-style model that captures many of these assumptions, and analyze sample-complexity issues in this setting: that is, how much of each type of data one should expect to need in order to learn well, and what are the basic quantities that these numbers depend on. The high-level idea of the proposed model is to augment the notion of a *concept class*, which is a set of functions (like linear separators or decision trees), with a notion of *compatibility* between a target function and the data distribution. That is, rather than talking of "learning a concept class $\mathcal{C}$," as one would do in the standard PAC model, we would talk of "learning a concept class $\mathcal{C}$ under compatibility notion χ." Furthermore, we require that the degree of compatibility be something that can be estimated from a finite sample: specifically, χ is actually a function from $\mathcal{C} \times X$ to $[0,1]$, where the compatibility of h with D is $\mathbf{E}_{x\in D}[\chi(h,x)]$. The degree of *in*compatibility is then something we can think of as a kind of "unlabeled error rate". This setup allows us to analyze the ability of a finite unlabeled sample to reduce our dependence on labeled examples, as a function of the compatibility of the target function and various measures of the helpfulness of the distribution. The type of implication our analysis provides is that unlabeled data is useful if (a) we have a good notion of compatibility so that the target function indeed has a low unlabeled error rate, (b) the distribution D is *helpful* in the sense that not too many other hypotheses also have a low unlabeled error rate, and (c) we have enough *unlabeled* data to estimate unlabeled error rates well.

We study a number of technical issues that arise in this context, such as what is the best way to describe the "complexity" of the set of functions with low unlabeled error rate, and how do the complexity of the hypothesis class $\mathcal{C}$ and compatibility notion χ combine to describe the number of unlabeled data points needed to provide uniformly good compatibility estimates. We provide sample-complexity bounds both for uniform convergence and specific ϵ-cover based algorithms.

Email address: ninamf@cs.cmu.edu

Empirical Bayesian test of the smoothness

Eduard Belitser and Farida Enikeeva, Utrecht University and EURANDOM

We observe independent Gaussian data $X=(X_i)_{i\in\mathbb{N}}$, where $X_i \sim N(\theta_i, n^{-1})$ and $\theta = (\theta_i)_{i\in\mathbb{N}} \in \ell_2$ is an unknown parameter.
Let $\left\{\Theta_\beta\right\}_{\beta\in\mathcal{B}} \subseteq \ell_2$ be a familyof Sobolev subspaces $\Theta_\beta = \left\{\theta : \sum_{i=1}^\infty i^{2\beta}\theta_i^2 < \infty\right\}$ with $\mathcal{B} \subseteq (1/2, +\infty)$ such that $\mathcal{B} = \{\beta_{-L_n}, \ldots, \beta_{-1}, \beta_0, \beta_1, \ldots, \beta_{K_n}\}$, where $\beta_i \leq \beta_{i+1}$ for all $\beta_i, \beta_{i+1} \in \mathcal{B}$. We suppose that θ belongs to a certain ellipsoid Θ_β for some $\beta \in \mathcal{B}$. Without loss of generality we can assume that $\theta \in \Theta_{\beta_0}$, otherwise we relabel the β_is.

Our main goal is to make an inference about the smoothness β of the underlying signal θ based on the data X. More generally, under appropriate conditions, we should be able to reject the hypothesis $\beta_\theta(B) < \beta_0$ if $\theta \in \Theta_{\beta_0}$. According to the Neyman-Pearson paradigm, we would like to test the hypothesis

$$H_0 : \ \beta_\theta(B) \geq \beta_0 \qquad \text{against} \qquad H_1 : \beta_\theta(B) \leq \beta_0 - \delta_n \,,$$

for some $\beta_0 \in \mathcal{B}$ and some positive sequence $\delta_n = \delta_n(\theta, \beta_0)$, which we woud like to tune as precisely as possible.

We construct a test statistic using the empirical Bayes approach. The idea of our approach is to put a "right" prior $\pi(\beta)$ on the parameter θ, find the marginal distribution of X which will depend on β, and then use the marginalized maximum likelihood estimator for β as the test statistic.

We put the Gaussian prior $\pi = \pi(\beta)$ on θ: the θ_i's are independent and for $\delta \in [0, 1]$

$$\theta_i \sim \mathcal{N}(0, \tau_i^2(\beta)) \,, \qquad \tau_i^2(\beta) = \tau_i^2(\beta, \delta, n) = n^{\frac{\delta-1}{2\beta+1}} i^{-(2\beta+\delta)} \,, \quad i \in \mathbb{N} \,.$$

Then the marginalized likelihood of the data X has the following form:

$$L_n(\beta) = \prod_{i=1}^{\infty} \frac{1}{\sqrt{2\pi\big(\tau_i^2(\beta) + n^{-1}\big)}} \exp\bigg\{ -\frac{X_i^2}{2\big(\tau_i^2(\beta) + n^{-1}\big)} \bigg\}.$$

For any set $S_n \subset \mathcal{B}$, define the marginalized maximum likelihood estimator $\widehat{\beta}(S_n) = \widehat{\beta}(S_n, X, n)$ of β restricted to the set S_n such that

$$L_n(\widehat{\beta}(S_n)) \geq L_n(\beta) \quad \text{for all } \beta \in S_n.$$

For a particular $\theta \in \ell_2$ define a function $A_\theta(\beta) = \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2$, $\beta \in \mathcal{B}$. The next theorem claims effectively that the type I error probability converges to zero.

Theorem 1. Suppose $A_\theta(\beta_0) < \infty$ for some $\beta_0 \in S_n \subset \mathcal{B}$, $|S_n^-(\beta_0)| \leq C_1 \exp\big\{C_2 n^{1/(2\beta_0+1)}\big\}$ for some $C_1, C_2 > 0$, and $\tau_i^2(\beta) = i^{-(2\beta+1)}$. Then there exist an integer $N = N(\beta_0, \theta, C_2)$ such that for any $C > 0$ there exists a positive $C_0 = C_0(\beta_0, C_2, C)$ such that for all $c > C_0$ the inequality

$$\mathbf{P}\Big\{\widehat{\beta}(S_n) \leq \beta_0 - \frac{c}{\log n}\Big\} < C_1 \exp\big\{ - C n^{1/(2\beta_0+1)}\big\}$$

holds for all $n \geq N$.

Email address: enikeeva@eurandom.tue.nl

Singularity, Instability, and the Variance-Bias Tradeoff

S. P. Ellis, Columbia University

Descriptive statistics, statistical inference, and learning all use "data analytic functions" (dafs) to extract structure from noisy data. We use mainly topological methods to study the stability of dafs. The approach is natural, however, explicit connection between our approach and probabilistic treatments of data analysis remains to be developed. (Because the results are topological they apply to perception and cognition by brains, not just to dafs defined by algorithms.)

The performance of a daf depends heavily on the daf's stability or conditioning: how much the daf's value changes in response to small changes in the data. We study the stability of dafs by studying the most extreme forms of instability, what I call "singularity." Let Φ be a daf defined on a dense subset $\mathcal{D}'$ of the "data space", $\mathcal{D}$. A singularity of Φ is a data set, $D \in \mathcal{D}$, at which the limit, $\lim_{D' \to D,\, D' \in \mathcal{D}'} \Phi(D')$, does not exist. So a singularity,

D, is like a discontinuity except that Φ does not have to be defined at D. An example is the well known "collinearity" phenomenon in least squares regression: If the matrix of predictors is not of full rank then tiny changes in the data can cause wild gyrations in the least squares fit. The set of all singularities of Φ is its "singular set," $\mathcal{S}$. *A crucial point is that* Φ *will be unstable near* $\mathcal{S}$. A working hypothesis is that, in fact, the singular set of a daf has a big impact on its stability.

A "test pattern" is a data set whose structure (of the kind that is of interest) is obvious and unambiguous. One expects a daf to perform well near an appropriate test pattern. E.g., if one is trying to fit a k-dimensional plane to multivariate data (call this activity "plane-fitting") an obvious test pattern is a data set that lies exactly on a unique k-plane. I dub failure of the daf to properly fit a data set in $\mathcal{T}$ "bias." The connection between this kind of bias and the standard notion of bias remains to be worked out, but I expect the two concepts to be related. Similarly, the behavior of a daf near its singular set contributes to its variance.

One would like the singular set of a daf to be small. Quite generally, if there are many singularities in $\mathcal{T}$ (large bias), then the singular set can have smaller Hausdorff dimension (less variance) than if there are few singularities in $\mathcal{T}$ (little bias). This is an example of a geometric interpretation of the "variance-bias tradeoff." In the absence of regularization we can use topological methods to compute tight general lower bounds on the the Hausdorff dimension of the singular sets of hypothesis tests, plane fitters, linear classifiers, and measures of location on spheres. We can also compute a crude general lower bound on the Hausdorff measure of the singular sets of many dafs. This bound provides more insight into the variance-bias tradeoff.

This theory applies indirectly to regularized methods. Let Λ be a regularized learning method. One can imbed the data space, $\mathcal{D}$, into a larger space, $\mathcal{E}$. On $\mathcal{E}$ there is a daf, Λ^*, whose restriction to a subspace, $\mathcal{E}_\Lambda$, is essentially Λ. The restriction of Λ^* to another subspace, $\mathcal{E}_0$, of $\mathcal{E}$ is an unregularized method. On part of $\mathcal{E}_0$ (call it $\mathcal{P}$) the daf Λ^* has little bias. It is tempting to reduce the bias of Λ by modifying it in order to bring the corresponding part of $\mathcal{E}_\Lambda$ close to $\mathcal{P}$. (Cross-validation will do this.) However, the theory described above then implies that some part of $\mathcal{E}_\Lambda$ will thereby be brought near a singularity of Λ^*. This will increase the instability, i.e., "variance," of Λ. This is another geometric interpretation of the "variance-bias" tradeoff.

Unit 42, NYSPI, 1051 Riverside Dr., New York, NY 10032, U.S.A..

Email address: ellis@neuron.cpmc.columbia.edu

Kernels, Associated Structures and Generalizations

Matthias Hein, Olivier Bousquet

Positive definite kernels are extremely powerful and versatile tools. They allow to construct spaces of functions on an arbitrary set with the convenient structure of a Hilbert space. Methods based on such kernels are usually very tractable because of the particular structure (reproducing property) of the space of functions. This has a large number of applications, in particular for statistical learning, approximation or interpolation where one has to manipulate functions defined on various types of data.
Our goal is to survey some of the results relevant for machine learning. Since the literature is scattered among various fields of mathematics we believe that the learning community would benefit from a unified exposition of the results and relationships between them. This work is a first attempt to go into that direction.
A first message is that there is an equivalence (in a strong) sense between several objects: positive definite (PD) kernels , reproducing kernel Hilbert spaces of functions, Gaussian processes and a class of positive operators. A second message is that the mysterious "feature maps" associated to PD kernels are not related to the Mercer property and they exist and can be defined in many different ways. A third message is that the integral operator associated to a PD kernel has nice properties even if the kernel is not continuous. In particular it is tightly related to the covariance operator (i.e. the population limit of a covariance matrix) as they have the same spectrum.
After these basic notions we consider three directions of generalization. One direction are operator-valued positive definite kernels, the second is the theory of Hilbertian subspaces introduced by Schwartz. We argue that these first two approaches end up being special cases of the basic theory. This may seem surprising but it can be easily seen by changing in both cases the index set one considers.

Finally the third direction of generalization is the theory of indefinite kernels (i.e. kernels that are not positive definite) and their associated structures, based on the notion of reproducing kernel Pontryagin respectively reproducing kernel Krein spaces.
Finally, we have to say that this work is far from being complete and there exist many other notions to be explored (and made accessible to the community). We hope to be able to provide an extended and more comprehensive account (covering for example Gaussian measures, generalized stochastic processes, group representations in RKHS, spectral decompositions of ker-

nels, regularization theory and various results of applications in approximation, interpolation etc.) in the near future.

Email address: mh@tuebingen.mpg.de

Relative loss bounds for predicting almost as well as any function in a union of Gaussian reproducing kernel spaces with varying widths

Mark Herbster, University College London

We consider a two-layer network algorithm. The first layer consists of an uncountable number of linear units. Each linear unit is an *LMS* algorithm whose inputs are first "kernelized." Each unit is indexed by the value of a parameter corresponding to a parameterized reproducing kernel, here an isotropic Gaussian Kernel parameterized by its width. The first-layer outputs are then connected to an *Exponential Weights* algorithm which combines them to produce the final output. We give performance guarantees for this algorithm.

As a guarantee of performance, we give a *relative loss bound* for this online algorithm. By online, we refer to the fact that learning proceeds in trials where on each trial the algorithm first receives a pattern, then it makes a prediction, after which it receives the true outcome, and finally incurs a loss on that trial measuring the discrepancy between its prediction and the true outcome. By relative loss bound, we refer to the fact on any trial, we can bound the cumulative of loss of the algorithm by the cumulative loss of any predictor in a comparison class of predictors plus an additive term. Hence the goal is that the performance of algorithm be almost as good as any predictor in the class; therefore we desire a small additive term. Often these bounds may be given without any probabilistic assumptions. In this note the comparison class is the set of functions obtained by a union of reproducing kernel spaces formed by isotropic Gaussian kernels of varying widths.

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK.

Email address: M.Herbster@cs.ucl.ac.uk

New Boosting Algorithms for Heteroskedastic Time Series

J. M. Matías[1], M. Febrero[2], W. González-Manteiga[2], and J. C. Reboredo[3]

This work develops and evaluates the following set of new algorithms based on boosting techniques, designed to model and predict heteroskedastic time series:

(i) Gradient_Boost algorithms with a loss function based on likelihood and which can simultaneously model both trend and volatility. The main novel elements of these algorithms are:

  (a) Its simultaneous application to trend and volatility components of the likelihood.

  (b) The use of likelihood-based models (e.g., ARMA-GARCH) as the base hypothesis rather than gradient fitting techniques using least squares, with a view to modelling any possible heteroskedasticity that the gradient series may inherit from the original series. In these conditions, the algorithm cannot be considered a genuine Gradient_Boost approach.

(ii) A new boosting algorithm (which we call Wild_BoostGARCH), specifically designed for the modelling of volatility of returns series. This algorithm uses the GARCH model as its base hypothesis and its philosophy is to gather in each iteration the amount of the variance that is not *explained* by the algorithm until that moment and introduce it again into the model (the GARCH of the next iteration) via a series of squared random Gaussian numbers whose variance is the amount of *unexplained* variance.

An evaluation of these algorithms with a single lag in various sets of artificial data and in the S&P 500 returns series produces the following conclusions:

(i) For all the sets of data analyzed, there was always at least one algorithm that significantly improved on the results of a GARCH model. On the other hand, no algorithm was revealed as the best in all the situations studied.

(ii) The Gradient_Boost algorithms tend to produce good results in all situations. It would appear that the empirical gradient is not merely a simple n-dimensional vector but a time series in itself, with possible heteroskedasticity inherited from the original series.

(iii) The Wild_BoostGARCH algorithm is potentially of interest when the most recent observations in the series suggest a possible future increase in volatility. With its additional injection of variance, the Wild_BoostGARCH algorithm has an effect similar to the assumption, in a GARCH model, of a Student-t distribution, whose heavier tails make the model postulate greater variance for the series.

[1]Dep. de Estadística, Universidad de Vigo, 36200 Vigo, Spain; [2]Dep. de Estadística, Universidad de Santiago de Compostela, Spain; [3]Dep. de Fundamentos del Análisis Económico, Universidad de Santiago de Compostela, Spain..

Email address: jmmatias@uvigo.es.

Kernel-based Fisher Discriminant Analysis for Microarray Data

Reverter, F. and Vegas, E., University of Barcelona.

A reliable and precise classification of tumors is essential for successful treatment of cancer. Current methods for classifying human malignancies rely on a variety of morphological, clinical, and molecular variables. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies which are being used increasingly in cancer research. By allowing the monitoring of expression levels for thousands of genes simultaneously, such techniques may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification.

In this paper, we consider recent machine learning approaches as kernel Fisher discriminant analysis for the classification of tumors based on gene expression profiles.

To evaluated the performance of the kernel Fisher discriminant analysis for microarray data we have used the leukemia dataset, available at `www.broad.mit.edu/cgi-bin/cancer/datasets.cgi` , and described in the recent paper of Golub *et al.* This dataset comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing 6817 human genes. The data comprise 47 cases of ALL and 25 cases of AML.

We have determined the kernel parameter and the regularization constant using *leave-one-out* cross validation. Results indicate that kernel-based approaches achieve competitive classification error rate for microarray data analysis.

Department of Statistics. University of Barcelona. Diagonal, 645. 08028 Barcelona. Spain.

Email address: freverter@ub.edu, evegas@ub.edu.

Interpolation of subspaces and the approximation error in learning theory

Peter Sunehag, Uppsala University

In "On the mathematical foundations of learning theory" by Cucker and Smale and in "Estimating the approximation error in learning theory" by Smale and Zhou the following problem was studied:
Let B be a Banach space (for example L^2) and let H be a dense subspace with $\|b\|_B \leq \|b\|_H$ for $b \in H$ (for example a Sobolev space $W^{s,2}$). H is called the Hypothesis space. Given $a \in B$ what is the convergence rate of the function

$$I(a,R) = \inf_{\|b\|_H \leq R} \|a-b\|, \ R > 0$$

as $R \to \infty$. To describe the space of all $b \in B$ which has a certain polynomial convergence rate they used interpolation theory. The interpolation spaces $(B,H)_{\theta,\infty}$ are defined from Peetre's K-functional,

$$K(t,a) = K(t,a,(B,H)) = \inf\{\|a_0\|_B + t\|a_1\|_H \mid a_0 + a_1 = a\}.$$

For $0 < \theta < 1$ the space $(B,H)_{\theta,\infty}$ is defined by the norm

$$\|a\|_{\theta,\infty} = \sup_{t>0} \frac{K(t,a)}{t^\theta}.$$

It is a known fact that

$$I(a,R) = O(R^{-\frac{\theta}{1-\theta}})$$

if and only if $a \in (B,H)_{\theta,\infty}$. The purpose of this presentation is to discuss the following question:
If we replace the Hypothesis space H with a closed subspace $\tilde{H} \subset H$ (in our $(L^2, W^{s,2})$ example we can for example define $\tilde{H}$ by adding a Neumann-type boundary condition) which is still dense in B, then for which $a \in B$ does

the convergence rate change. More precisely, in interpolation terminology, the question that will be discussed is:
For which θ does it hold that $(B,\tilde{H})_{\theta,\infty} = (B,H)_{\theta,\infty}$ with equivalent norms? That kind of problems are usually refered to as interpolation of subspaces. The author of this abstract has proved that there is an indice γ such that $(B,\tilde{H})_{\theta,\infty} = (B,H)_{\theta,\infty}$ when $\theta < \gamma$. It is defined by letting

$$\gamma = \lim_{k\to-\infty} \inf_n \frac{1}{k} \log_2 \frac{\kappa(2^{n+k})}{\kappa(2^n)}$$

where

$$\kappa(t) = \sup_{f\in\tilde{H}^{\perp},\ \|f\|_{H'}\le 1} K(t,f,(B',H')).$$

Department of mathematics, Uppsala University, Box 480, SE-751 06 Uppsala, Sweden.

Email address: peter@math.uu.se

Error Analysis for Linear Programming SVM Classification

Qiang Wu and Ding-Xuan Zhou, Department of Mathematics, City University of Hong Kong

In this paper, we consider the linear programming (LP) support vector machine (SVM) classification with Mercer kernels and provide error analysis. As the 1–norm soft margin SVM has been well understood, we analyze the LP–SVM by setting a stepping stone between the LP–SVM and the 1–norm soft margin SVM. An upper bound for the misclassification error is obtained. The rate analysis based on this upper bound shows that the learning rates of the LP–SVM and the 1–norm soft margin SVM are the same for deterministic distributions. As for distributions with noise, they are still the same if the capacity of hypothesis space is small while the learning rate of the LP–SVM is slightly worse otherwise.

We also consider the LP–SVM with kernels which are not necessarily positive definite and obtain some upper bounds. However, things are more complicated and many problems keep open for this algorithm.

Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China.

Email address: wu.qiang@student.cityu.edu.hk, mazhou@cityu.edu.hk

Regularization error in SVM classification

Yiming Ying, Department of Mathematics, City University of Hong Kong, China

The purpose of this paper is to estimate the regularization error for $q-$norm soft margin SVM classification algorithm. First, We observe that the regularization error can be bounded by a $\mathcal{K}-$functional for L^q spaces in approximation theory. This enables us to give general convergence rates for the regularization error of $q-$norm SVM under some assumptions. Examples for Gaussian kernel and spline kernel are given to demonstrate this general result. Secondly, Faster rates are obtained if the variance σ of the Gaussian kernel K_σ is flexible (while the union of balls in $\mathcal{H}_{K_\sigma}$ with a uniform radius 1 is a UGC class). At Last, the condition of weakly separable distribution is introduced which is sufficient for the polynomial rate of regularization error. This is joint work with Prof. Chen ,Q.Wu and Dr. D.X. Zhou.

Institute of Mathematics, Chinese Academy of Science, Beijing 100080, CHINA..

Email address: ymying@cityu.edu.hk

5 List of participants

Abdallah, Mounif Hassan	mabdallah@qf.org.qa
Alquier, Pierre	alquier@ensae.fr
Audibert, Jean-Yves	audibert@ureach.com
Balcan, Maria-Florina	ninamf@cs.cmu.edu
Barron, Andrew andrew.	barron@yale.edu
Barthe, Franck	barthe@math.ups-tlse.fr
Bartlett, Peter L.	bartlett@cs.berkeley.edu
Belouardy, Nabil	nabil.belouardy@eurocontrol.int
Boistard, Hélène	helene@eio.uva.es
Boucheron, Stéphane	stephane.boucheron@lri.fr
Castro, Rui	rcastro@rice.edu
Cesa-Bianchi, Nicolò	cesa-bianchi@dsi.unimi.it
Chacón, José Enrique	jechacon@eio.uva.es
Chen, Zhe	zhechen@soma.crl.mcmaster.ca
Clémençon, Stéphan	sclemenc@u-paris10.fr
Dahmani, Zoubir	dahmanizz@yahoo.fr
De Bie, Tijl	tijl.debie@esat.kuleuven.ac.be
Dekel, Ofer	oferd@cs.huji.ac.il
DeVore, Ronald	devore@math.sc.edu
Devroye, Luc	luc@cs.mcgill.ca
Dudley, Richard M.	rmd@math.mit.edu
Ellis, Steven	ellis@neuron.cpme.columbia.edu
Enikeeva, Farida	enikeeva@eurandom.tue.nl
Esposito, Yann	esposito@cmi.univ-mrs.fr
Fernandez Temprano, Miguel Alejandro	miguellaf@eio.uva.es
Foster, Dean	dean@foster.net
Fradelizi, Matthieu	matthieu.fradelizi@univ-mlv.fr
Freund, Yoav	freund@cs.columbia.edu
Gey, Servane	Servane.Gey@math-info.univ-paris5.fr

Giné, Evarist	gine@math.oconn.edu
Giraud, Christophe	cgiraud@math.unice.fr
Guedon, Olivier	guedon@math.jussieu.fr
Hart, Sergiu	hart@huji.ac.il
Hein, Matthias	mh@tuebingen.mpg.de
Kääriäinen, Matti	matti.kaariainen@cs.helsinki.fi
Kirik, Ekaterina	kirik@icm.krasn.ru
Koltchinskii, Vladimir	vlad@math.unm.edu
Kondor, Imre Risi	risi@cs.columbia.edu
Krämer, Nicole	nkraemer@cs.tu-berlin.de
Kurková, Vera	vera@cs.cas.cz
Ledoux, Michel	ledoux@math.ups-tlse.fr
Leung, Gilbert	gilbert.leung@yale.edu
Linial, Nati	nati@cs.huji.ac.il
Litvak, Alexander	alexandr@math.ualberta.ca
Livchits, Evgueni	livshitz@mccme.ru
Long, Phil	plong@cs.columbia.edu
Lugosi, Gábor	lugosi@upf.es
Mary-Huard, Tristan	maryhuar@inapg.fr
Massart, Pascal	Pascal.Massart@math.u-psud.fr
Matías, José María	jmmatias@uvigo.es
Mendelson, Shahar	shahar.mendelson@anu.edu.au
Milman, Vitali	milman@post.tau.ac.il
Naish-Guzman, Andrew	agpn@cam.ac.uk
Naor, Assaf	anaor@microsoft.com
Neumeyer, Natalie	natalie.neumeyer@rub.de
Nishimori, Yasunori	y.nishimori@aist.go.jp
Opper, Manfred	m.opper@aston.ac.uk
Pajor, Alain	alain.pajor@univ-mlv.fr
Paquet, Ulrich	up208@cam.ac.uk
Philips, Petra	petra.philips@anu.edu.au

Pinkus, Allan	pinkus@math.technion.ac.il
Rafikov, Evgeniy	erafikov@mccme.ru
Rakhlin, Alexander	rakhlin@mit.edu
Reverter, Ferran	freverter@ub.edu
Rigollet, Philippe	rigollet@ccr.jussieu.fr
Rivieccio, Fabio	rivieccio@dibe.unige.it
Rosasco, Lorenzo Andrea	rosasco@disi.unige.it
Samson, Paul-Marie	samson@math.univ-mlv.fr
Schechtman, Gideon	gideon@weizmann.ac.il
Schreibman, Adi	adidan@cs.huji.ac.il
Scott, Alena	oetting@stat.rice.edu
Scott, Clayton	cscott@rice.edu
Shah, Parantu	shah@embl.de
Smale, Stephen	smale@math.berkeley.edu
Stoltz, Gilles	gilles.stoltz@math.u-psud.fr
Sunehag, Peter	peter@math.uu.se
Szepesvari, Csaba	szcsaba@sztaki.hu
Tarres, Pierre	tarres@cict.fr
Tomczak-Jaegermann, Nicole	nicole.tomczak@ualberta.ca
Tsybakov, Alexandre	tsybakov@ccr.jussieu.fr
Utzet, Frederic	utzet@mat.uab.es
Vayatis, Nicolas	vayatis@ccr.jussieu.fr
Vegas, Esteban	evegas@ub.edu
Vershynin, Roman	vershynin@math.ucdavis.edu
Vovk, Vladimir	vovk@cs.rhul.ac.uk
Wellner, Jon	jaw@stat.washington.edu
Wu, Qiang	wu.qiang@student.cityu.edu.hk
Ying, Yiming	ymying@cityu.edu.hk
Zaiats, Vladimir	zaiats@mat.uab.es
Zhou, Ding-Xuan	mazhou@math.cityu.edu.hk
Zinn, Joel	jzinn@math.tamu.edu
Zwald, Laurent	laurent.zwald@math.u-psud.fr