Inference Principles and Model Selection

23.07 - 27.07.2001

organized by

Joachim Buhmann, Bernhard Schölkopf

The core problem of statistics and machine learning addresses the question how can we efficiently find a statistical model to describe empirical data. Classical statistical approaches to solve this problem have been complemented during the last 15 years by Neural Computation, a very promising strategy to data analysis. The Dagstuhl seminar on "*Inference Principles and Model Selection*" — the fourth in a series of Machine Learning and Neural Computation workshops in 1994, 1997, 1999 and 2001 — was intended to review this exciting development of the field and to discuss the foundation of statistical and computational learning theory with its deep (and still unresolved) questions. The participants represented all of the involved disciplines from statistics and computer science to information theory and philosophy. The burning question of many participants if there exist notions of inference studied in philosophy that machine learning has overlooked so far came up in several sessions and especially in the first tutorial on Philosophical Foundations (Matthias Hild). Three pioneers of the field, Sun-ichi Amari, Phil Dawid and Vladimir Vapnik provided valuable insights how the field of learning machines developed from the sixties up to today and what kind of challenges are lying still ahead of us.

What have been the main conclusions of the seminar?
In contrast to the previous seminars, this workshop with its tutorials and short position statements (rather than conference style talks) forced the participants to concentrate on conceptual issues with as little obstruction as possible by technical details. Common ground between Bayesian inference, statistical and computational learning theory and logical approaches to inference as well as concepts from information theory have been observed and widely discussed.

The final discussion session summarized the following open issues of the field:

1. Tali Tishby reminded us that learning and information extraction goes beyond the issues of sample fluctuations which are extensively studied in the Computational Learning Theory community. What are the correct inference principles to detect structures hidden in data?

2. How can we evaluate learning principles and algorithms? How should we design good experiments?

3. Is model selection or model combination more effective in structure detection?

4. How can we find more characterization results for learning algorithms? What is an appropriate size of the validation set?

5. How should we proceed in non i.i.d. situations where data are dependent? How can the concepts from classification and regression be extended to time series analysis and to Markov random fields.

6. What is the correct number of inference levels?

Most of these questions will stay with us for the next decades but this workshop has raised the awareness of all participants which parts of machine learning and neural computation are based on fundamental principles and where we still have to discover such a solid foundation.

Bonn, 30. Juli 2001 Joachim M. Buhmann

Remark:
These abstracts and links to slides are available at `www-dbv.cs.uni-bonn.de/dagstuhl01`.

Contents

1 Information Geometry and Inference Principles

Shun-ichi Amari
RIKEN Brain Science Institute

Information geometry studies the intrinsic geometrical structure of a family of probability distributions. The structure is uniquely defined from the principle of invariance, giving a Riemannian metric (due to the Fisher information matrix) and a dual pair of affine connections. It is useful in many problems related to stochastic phenomena such as statistical inference, model selection, information theory, control systems theory, etc.

We apply the method of information geometry to multilayer perceptrons, which have nonlinear input-output relations depending on the modifiable parameters. They are modified by learning from examples. When noises disturb the output, the behavior of a multilayer perceptron is described by the conditional probability distribution of the output conditioned on the input, and the probability distribution is parameterized by the modifiable parameters.

The parameter space, called the neuromanifold, is a family of probability distributions in which the learning process is represented by a trajectory. The stochastic gradient learning method is most popular in on-line learning. However, when the parameter space has a Riemannian structure, the gradient does not represent the true steepest direction and should be replaced by the Riemannian or natural gradient. The backprop method is notorious for slow convergence, due to plateaus. Such plateaus are created by the underlying geometrical structure, and the natural gradient method is shown to have a very good convergence property. It is, however, difficult to calculate the Fisher information matrix explicitly and to invert it. We give an adaptive method of obtaining the inverse of the Fisher information matrix directly.

If the neuromanifold is not so strongly curved, the natural gradient is not so different from the ordinary gradient. This suggests that the neuromanifold is strongly curved. We show many hierarchical structures such as multilayer perceptrons, Gaussian mixtures, ARMA models in time series, etc., include singular points in the parameter spaces, where the Fisher information matrix degenerates. The singularities are given rise to by its inner symmetry, and occurs at the points on which the system parameters become redundant. Model selection is important when the true system lies in a neighborhood of such singularities.

Therefore, we need to analyze the behaviors of statistical inference and learning, when the true system lies in a neighborhood of a singular point. The conventional Cramer-Rao paradigm does not hold in such a case, because the Fisher information is degenerate. The central limit theorem cannot be applied, either.

One should remark that model selection is important in such a situation, but the conventional theories of AIC and MDL are based on the Cramer-Rao paradigm which does not hold. Hence, we need to have a new theoretical paradigm. The Baysian framework should be also modified.
The present talk will discuss these aspects of geometry of neuro- manifolds in connection with learning, inference and model selection.

2 Gradient Estimates in Reinforcement Learning

Peter Bartlett
BIOwulf Technologies, Inc.

We consider the problem of controlling a partially observable Markov decision process (POMDP), so as to maximize the time average of a reward criterion. For parameterized stochastic policies, one approach is to use the gradient of the performance criterion with respect to the policy parameters. We present algorithms to estimate these gradients from a single sample path, by relying on mixing properties of the controlled POMDP. We give bounds on the estimation and approximation errors fo these estimates for finite samples, in terms of a certain mixing time of the controlled POMDP. The variance of these Monte Carlos estimates can be reduced using additive control variate methods. Two commonly used approaches, reward baselines and actor-critic algorithms, are special cases. We present bounds on the expected error for these algorithms, and derive the baselines and critics that minimize these bounds. These results allow us to evaluate how suboptimal commonly used algorithms are, and lead to new algorithms for gradient estimates.
(joint work with Evan Greensmith and Jonathan Baxter)

3 Concentration inequalities and penalization methods in model selection

Stephane Boucheron
Laboratorie de Recherche en Informatique,
CNRS - Université Paris-Sud (FR)

Concentration inequalities constitute "natural" extensions of the classical exponential bounds for sums fo independent random variables. (Azuma-Hoeffdings, Bennett, Bernstein). Concentration may be regarded as a "new look at independence". The basic message may be formulated as follows: any function of many independent random variables that is smooth in an appropriate sense is almost constant. The definition of smoothness or alternatively of enlargement of sets, starting from smoothness w.r.t. Hamming distance, to the recent formulations by Talagrand, is not straightforward. Such extensions are very useful when trying to characterize the fluctuations of quantities such as empirical VC-dimension, empricial VC-entropies, Rademacher complexities. Those last results can be obtained using the relatively transparent "entropy method" proposed by Ledoux - One of the killer applications of the concentration approach was the tails of suprema of empirical processes indexed by bounded functions (Talagrand 96, Ledoux 97, Massart 2000, Rio 2001) Concentration inequalities deal with the very topic of the Vapnik-Chervonenkis inequalities. In contrast to the latter, concentration inequalities only deal with fluctations around the mean, leaving the characterization of this mean to chaining techniques (for empirical processes).
As far as model selection is concerned, concentration inequalities prove useful in the design and anlysis of penalization strategies as advocated by Birgé and Massart, Vapnik, etc. In the Structural Risk Minimization framework, we are interested in selecting among a set of models $F_1, \ldots, F_k, \ldots$, i.e. among a set of estimates $\hat{f}_1, \ldots, \hat{f}_k, \ldots$ obtained from some randomly collected data set $D_n = ((x_1, y_1), \ldots, (x_n, y_n))$ such that the effective loss $E[l(\hat{f}(X), Y)]$ is minimal. (l might be absolute, quadratic loss)
The basic idea is that one should minimize a penalized empirical risk

$$L_n(\hat{f}_k) = \sum_{i=1}^{n} \frac{1}{n} l(\hat{f}_k(x_i), y_i)$$

plus some penalty term $pen(n, k)$. $pen(n, k)$ should minimize the amount of overfitting in F_k, i.e. $L(\hat{f}_k - L_n(\hat{f}_k)$.
Concentration inequalities prove useful in designing and analyzing *data-dependent* penalties. The latter constitute an important ingredient in any would-be practical system with guaranteed performance.

It remains the question to determine when such data-dependent penalization techniques can achieve adaptivity in the Donoho-Johnstone sense.

4 Tracking a Small Set of Experts by Mixing Past Posteriors

Olivier Bousquet
École Polytechnique, Centre de Mathématiques Appliqués,
Palaiseau (FR)

[slides available electronically, see preface]We examine on-line learning problems

in which the target concept is allowed to change over time. In each trial a master algorithm receives predictions from a large set of n experts. Its goal is to predict almost as well as the best sequence of such experts chosen off-line by partitioning the training sequence into $k+1$ sections and then choosing the best expert for each section. We build on methods developed by Herbster and Warmuth and consider an open problem posed by Freund where the experts in the best partition are from a small pool of size m. Since $k >> m$ the best expert shifts back and forth between the experts of the small pool. We propose algorithms that solve this open problem by mixing the past posteriors maintained by the master algorithm. We relate the number of bits needed for encoding the best partition to the loss bounds of the algorithms. Instead of paying $\log n$ for choosing the best expert in each section we first pay $\log \binom{n}{m}$ bits in the bounds for identifying the pool of m experts and then $\log m$ bits per new section.
(joint work with M.Warmuth)

5 Learning and Combinatorial Optimization: The noisy Traveling Salesman Problem

Joachim Buhmann
Institut für Informatik, Universität Bonn

[Abstract available electronically, see preface]

Many problems in the real world which are modeled as combinatorial optimization problems are stochastic in nature, i.e. the parameters defining the problem are random variables. This fact is traditionally neglected when a combinatorial optimization problem is formulated. I demonstrate with an example of the traveling salesman, that the minimal solution computed on a single (training) instance of a random problem can perform suboptimally on a second (test) instance. Computer experiments provide empirical evidence that certain Markov Chain Monte Carlo algorithms yield solutions which are more robust than the optimal training solution. Learning is performed by sampling a typical permutation matrix or by suitably averaging over TSP solutions.
The overfitting behavior of the ERM solution can be understood in terms of statistical learning theory. The MCMC algorithm computes an approximation to the empirical risk and the approximation accuracy should be controlled by robustness against overfitting. Too precise approximations of the training risk overfit a test instance, whereas too crude approximations introduce an underfitting bias. A generalization of the VC inequality quantifies this bias variance tradeoff. Large deviations between training and test performance are bounded by Bernstein's inequality. The minimum of this bound determines the stop temperature for an annealing scheme.
(joint work with Mikio Braun)

6 Hilbertian Learning

Stephane Canu
INSA de Rouen, France

Kernels and in particular Mercer or reproducing kernels play crucial role in the statistical learning theory and functional estimation. But very few is known about the underlying functional space where algorithms are looking for the solution. How to choose it? How to build it? What is its relationship with regularization? Introducing *Hilbert-Schmidt* operators helps to answer some of these questions. This allow to introduce learnable *frames* as a powerful and promising functional tool to build relevant kernels. Furthermore the learnable *frames* framework clarify the relationship between kernels and parametric components. This is a theory for semi-parametric learning. In particular wavelets are included in this framework

together with their associated kernels.

7 Vicinal Risk Minimization

Olivier Chapelle
BIOwulf Technologies, Inc., Paris (FR)

The Vicinal Risk Minimization principle establishes a bridge between generative models and methods derived from the Structural Risk Minimization Principle such as Support Vector Machines or Statistical Regularization. We explain how VRM provides a framework which integrates a number of existing algorithms, such as Parzen windows, Support Vector Machines, Ridge Regression, Constrained Logistic Classifiers and Tangent-Prop.
We will show how the approach implies new algorithms for solving problems usually associated with generative models. New algorithms are described for dealing with pattern recognition problems with very different pattern distributions and dealing with unlabeled data.

8 Bayesian and Prequential Inference for Model Selection

A. P. Dawid
University College London, Dept. of Statistical Science, London (GB)

[Abstract available electronically, see preface]

The Bayesian approach to inference allows us to express, in simple probabilistic form, two kinds of uncertainty: both about an unknown parameter of an assumed model, and about the model itself. It also very naturally allows us to make predictions for as yet unobserved quantities, a feature that turns out to be particularly valuable for model selection, as well as important in its own right.
In my talk I shall describe how the Bayesian approach naturally behaves in an

asymptotically desirable way in problems of model selection, without the need for any extraneous or ad hoc ingredients such as regularisation, nor oversimplifying assumptions such as independent observations. Some of the problems arising with finite data-sets will also be considered.
I shall then generalise the Bayesian predictive approach by introducing the methodology of Prequential Analysis, which assesses and compares model directly in terms of their 1-step ahead predictive performance. It is thus naturally suited to the task of model criticism and model selection. An important result is the availability of "optimal" forecasts, based on an extension of Emprical Risk Minimisation, even when the model is incorrectly specified.

9 Model Selection and Infinite Models

Zoubin Ghahramani
University College London,
Gatsby Computational Neuroscience Unit, London (GB)

[Abstract available electronically, see preface]

I will discuss two apparently conflicting views of Bayesian Learning. The first invokes automatic Occam's Razor (which results from averaging over the parameters) to do model selection, usually preferring models of low complexity. The second advocates not limiting the number of parameters in the model and doing inference in the limit of a large number of parameters if computationally possible. The first view lends itself to methods of approximating the evidence such as variational approximations. I will briefly describe these and give examples. For the second view, I will show that for a variety of models it is possible to do efficient inference even with an infinite number of parameters. I will discuss pros and cons of both views and how they can be reconciled.
(Joint work with Carl E. Rasmussen and Matthew J. Beal.)

10 Bagging equalizes influence

Yves Grandvalet
Université de Technologie de Compiegne,

Dept. Genie Informatique, Compiegne (FR)

Bagging constructs an estimator by averaging predictors trained on bootstrap samples. Bagged estimates almost consistently improve on the original predictor. It is thus important to understand the reasons for this success, and also for the occasional failures. It is widely believed that bagging is effective thanks to the variance reduction stemming from averaging predictors. However, seven years from its introduction, bagging is still not fully understood.
We provide experimental evidence supporting that bagging stabilizes prediction by equalizing the influence of training examples. Bagging's improvements/deteriorations can be explained by the goodness/badness of highly influential examples, whereas other arguments reach their limits. Finally, the reasons for the equalization effect support that other resampling strategies such as half-sampling should provide qualitatively identical effects while being computationally more efficient than bootstrap sampling.

11 Kernel Methods

Isabelle Guyon
ClopiNet, Berkeley (USA)

Kernel methods address a wide variety of induction problems, including function approximation (interpolation and regression), classification, density estimation, clustering and solving linear operator equations. The history of kernel machines started in the 19th century when Hilbert and Schmidt introduced integral equations of the form

$$\int K(s,r)f(t)dt = F(s).$$

This triggered a lot of research on the conditions that the kernel function $K(s,t)$ must satisfy. In 1909 Mercer stated the equivalence of positive definite kernels and valid "dot products" that opened the doors to a lot of theoretical derivations. Kernels then appeared in density estimation (Parzen Windows), classifiction and regression (splines). Parzen windows type kernels are shift invariant and include Gaussian kernels and potential functions. They were introduced in the 1960's (Parzen 1962, Aizerman et al. 1964). Kernels are also similarity measures and include various dot products. One of the most widely used one is the polynomial kernel $(x \cdot y)^q$, q being the polynomial degree. Kernels have been used in signal processing (convolutions) and image processing. Using kernels in the preprocess-

ing stes leads to a a nice unified framework of creating new kernels by combining kernels. The duality between approximation functions linear in their parameters $f(x) = w \cdot \phi(x)$ and kernel approximationn functions

$$f(x) = \sum_k \alpha_k K(x, x_k)$$

is known since the 1960's. It has known a regain of interest since 1992 when it was first used in the context of support vector machines (SVMs). Since then a lot of algorithms that exploit this duality have been derived. Many extensions of the original simple kernel machines have been made allowing users to treat non-vectorial inputs (strings, tree, sets) and fancy outputs (multiclass, mulitlabel, sequences) and to address a variety of complex optimization problems. As of today kernel machines span a wide range of applications with a wide spectrum of sizes of input space and training data sets. The most popular kernels are the Guassian kernel and the polynomial kernel (with its special case the linear kernel) but specialized kernels are an active area of reserach.

12 Algorithmic Luckiness

Ralf Herbrich
Microsoft Research, Cambridge (GB)

[Abstract available electronically, see preface]

In contrast to standard statistical learning theory which studies uniform bounds on the expected error we present a framework that exploits the specific learning algorithm used. Motivated by the luckiness framework [Taylor et al., 1998] we are also able to exploit the serendipity of the training sample. The main difference to previous approaches lies in the complexity measure; rather than covering all hypotheses in a given hypothesis space it is only necessary to cover the functions which could have been learned using the fixed learning algorithm. We show how the resulting framework relates to the VC, luckiness and compression frameworks. Finally, we present an application of this framework to the maximum margin algorithm for linear classifiers which results in a bound that exploits both the margin and the distribution of the data in feature space.
(Joint work with Bob Williamson.)

13 Inductive Reasoning

Matthias Hild
California Institute of Technology & Jet Propulsion Laboratory, Pasadena

The paper reviews some of the problems of inductive reasoning that have been discussed in the philosophical literature. The presentation is intended for an interdisciplinary audience.

14 Gaps and Bridges between Inductive Inference and Statistical Learning Theory

Wolfram Menzel
Institute for Logic, Complexity and Deduction Systems
Computer Science Department, University of Karlsruhe (GE)

[Abstract available electronically, see preface]

These two worlds look totally different, hardly comparable to each other. Still, both come from a common intuition to model phenomena of learning. Analysis of their differences and possible relationship leads (among others) to three main points:

- Probability measures on the power set of N are "finitary", thus contrary to all "Lebesque-style" ones.
- Uniformity as commonly required in learnability definitions of statistical learning theory is a sensitive and crucial point.
- When a hypothesis space is parameterized, the relationship between distance measuring among parameters on the one hand, and among the meant functions on the other hand seems to be important, at least in application oriented approaches.

Results can be presented on two kinds of questions:

1. A tried combination of learning as "approximating in a uniform and arbitrarily good way" and as "finding an ultimate, stable regularity".

2. In the Inductive Inference scenario: Can hopes for some kind of "continuity" or at least "compatibility" between distance measuring among programs (parameters) and among computable functions ever be satisfied?

15 Assessing Reliability of Unsupervised Learning: a Resampling Approach

Klaus-Robert Müller
Frauenhofer Gesellschaft FIRST, Berlin and
Univ. of Potsdam, Am neuen Palais 10, Potsdam

When applying unsupervised learning techniques like ICA or temporal decorrelation, a key question is whether the discovered projections are reliable. In other words, can we give *error bars* or can we assess the *quality* of our separation? We use resampling methods to tackle these questions and show experimentally that our proposed variance estimations are strongly correlated to the separation error. We demonstrate that this reliability estimation can be used to select the appropriate ICA-model to enhance significantly the separation performance, and, most important, to mark the components that can really have a physical meaning. Application to data from an MEG experiment underlines the usefulness of our approach.
(Joint work with Frank Meinecke, Andreas Ziehe and Motoaki Kawanabe.)

16 Bias of Estimators and Regularization Terms

Noboru Murata
Department of Electrical, Electronics, and Computer Engineering,
Waseda University, Tokyo (JP)

[Abstract available electronically, see preface]

We deal with the role of regularization terms (penalty terms) from the view point of bias of the minimum training error estimation. In the field of neural

networks, for instance, regularization terms are often utilized to avoid over-fitting, however most of the time cross-validation is chosen to determine the strength of the regularization.
First we will clarify the bias of minimum training error estimation, which is caused by the nonlinearity of the learning system and depends on the size of training samples. Then taking this bias into account, we consider an appropriate size of the regularization term which is minimizing the predictive errors. The optimal size of the reguralization term in this sense is calculated from the second and third order information of the loss function. When the learning system has a large number of modifiable parameters, it is computationally expensive to calculate the higher order information, thus we propose a simple method of approximating the optimal size via a generalized AIC.

17 Models in Hyperbolic Space

Helge Ritter
Neuroinformatics Group, University of Bielefeld (GE)

A crucial question for model selection is the proper structural bias for the task at hand. Hyperbolic spaces offer in certain cases an attractive alternative to the usually employed Euclidean R^n, since they offer exponentially growing neighborhoods already in $d = 2$. After a brief synopsis of recent work we present as an example the use of the discretized hyperbolic plane for the creation of dimension-reduced mappings of text-document data, exhibiting semantic relationships as neighborhood in hyperbolic 2d-space.

18 Optimal Inductive Inference Under an Algorithmic Prior Reflecting Maximally Efficient Data Generation

Jürgen Schmidhuber
IDSIA, Lugano

Solomonoff's optimal but noncomputable strategy for inductive inference assumes the observations are drawn from a recursive prior. Here we make the additional assumption that the process computing the data is optimally efficient, and that the cumulative prior probability of all data whose computation costs at least $O(n)$ time is inversely proportional to n. Since in fact there exists a very simple, general, asymptotically optimal algorithm for *all* computable data, we can explicitly extract the corresponding *speed prior*, and derive a *computable* strategy for optimal inductive reasoning.

19 Stability of Posterior Estimates for Kernels

Alex Smola
Australian National University, Machine Learning Group, Canberra (AU)

Maximum a posteriori approximation is a popular technique for Gaussian Processes, since the value of the negative log-posterior can be used as an indicator of how plausible a certain hypothesis happens to be. The minimum of the regularized risk functionals in Support Vector Machines can be used for a similar purpose.
We prove that the minimum of the negative log-posterior and the regularized risk functional are concentrated random variables, provided the likelihood (or loss function) is a log-concave function.

20 Statistical Inference and Relevant Information Encoding

Naftali Tishby
The Hebrew University of Jerusalem, School of Computer Science & Engineering (IL)

The "Information bottleneck method" is an unsupervised non-parametric data organization technique that aims at extracting the relevant information in one

random variable with respect to another one. Given a joint distribution, $p(x, y)$, this method constructs a new variable $\hat{X}$ (or T) that infers partitions (soft) over the values of X that are informative about Y.
Many problems can be cast into this general framework, such as: time series prediction, supervised and unsupervised learning, noise filtering, feature extraction, etc. It can be formulated as a tradeoff between two mutual information measures:

$$L[p(\hat{x}|x)] = I(X; \hat{X}) - \beta I(\hat{X}; Y)$$

which as a closed set of self-consistent equations has to be satisfied at the stationary points of this langrangian for every value of the positive Lagrange multiplier β. We have proved the general convergence of an iterative algorithm - similar to the Blahut Arimoto algorithm in Rate-Distortion theory - that finds the optimal tradeoff and partition $p(\hat{x}|x)$. The algorithm has an agglomerative greedy version which has been applied successfully to problems such as document classification, gene expression analysis, spectral analysis and neuronal coding. We have recently extended the method to multivariate cases, by using the Lagrangian

$$L = I^{G_1}(X_1, \ldots, X_n, T) - \beta I^{G_2}(X_1, \ldots, X_n, T)$$

where $I^{G_{1,2}}$ are multiinformations of a Bayesian net.

21 Optimal aggregation of classifiers in statistical learning

Alexandre Tsybakov
Université Paris VI, URA - CNRS (FR)

The problem of statistical learning can be considered as a problem of nonparametric estimation of sets where the risk is defined by means of a specific distance function between sets associated to the misclassification error. The rates of convergence of classifiers depend on two parameters: the complexity of the class of candidate sets and the "margin" parameter. The dependence is explicitly given, in particular the optimal rates up to $O(n^{-1})$ can be attained where n is the sample size, and the proposed classifiers have the property of robustness to the margin. The main result of the paper concerns optimal aggregation of classifiers: we suggest a classifier that automatically adapts both to the complexity and to the margin, and attains the optimal fast rates, up to a logarithmic factor.

22 Development of Statistical Learning Theory

Vladimir Vapnik
AT&T Labs / Holloway College London (GB)

The development of Statistical Learning Theory is considered from the point of view of foundations of statistics. Two different foundatiaons of classical statistics are considered: the Glivenko-Cantelli-Kolmogorov (theoretical) approach, and Fisher's (simplified) applied approach. In the early 60s, it was realized that Fisher's approach is not sufficiently powerful for solving high-dimensional problems. Therefore, a theory continuing the ideas of the theoretical approach (Kolmogorov appraoch) was developed. Initially, it was rejected by the statistics community, which is why it found its home in computer science. Now, it is a well-developed branch of science studied by both statistics and computer sciences.

23 Predictive complexity: theory, possible applications, and open problems

Volodya Vovk
Royal Holloway, University of London (GB)

This talk will give a high-level review of some applications of Kolmogorov's notion of complexity and its variants to the problems of inference and model selection. We will argue that approaches to inference can be broadly classified as belonging to either "Bayesian" or "Popperian" paradigm. Kolmogorov complexity and its generalization, predictive complexity, are technical tools useful in both paradigms. In particular, we will discuss the use of Kolmogorov and predictive complexity in the MDL principle and its generalization, "Complexity Approximation Principle"; the latter will be contrasted with the Bayesian-type approach of the theory of prediction with expert advice. The notion of predictive complexity makes it possible to generalize the usual notions of randomness and information to a wide class of loss functions; this generalization allows us to formalize interesting questions about the limits of inference. Several open problems about predictive complexity will also be stated.

24 On-line learning - Methods and Open Problem

Manfred Warmuth
University of California, Santa Cruz, Dept. of Computer Science (USA)

[Abstract available electronically, see preface]

An "on-line" learning algorithm sees the examples one at a time and incurs a loss on each new example based on its current hypothesis. This hypothesis is updated on-line as more examples are seen. We are given a comparison class of predictors. The loss of the on-line algorithm on a sequence of examples is typically larger than the loss of the best off-line predictor in the comparison class. The goal of the learner is to bound the additional loss of the on-line algorithm over the best off-line predictor on an arbitrary sequence of examples. Such bounds are called "relative loss bounds" and quantify the price of hiding the future examples from the learner.
We discuss method for deriving on-line algorithms and for proving relative loss bounds. No background is required. We will stay at a high level and discuss directions for future research.
The key tool we use is Bregman divergences. They are used as loss functions and as measures of "distance" between two members of the comparison class.
We discuss families of algorithms that are characterized by different Bregman divergences. The two main families are the gradient descent and exponentiated gradient family. The former family includes all the kernel based algorithms and the latter family is motived by the minimum relative entropy principle (i.e. information theoretic motivation). We contrast the merits of both families of algorithm.

25 Reinforcement Learning with Many Parameters

Chris Watkins
Royal Holloway, Dept. of Computer Science, University of London (GB)

In my talk, I described a (well-known) reinforcement learning algorithm, the "Relative Payoff Process", which can be motivated both as a simple and biologically feasible learning method, and as a simple model of evolution. The performance of the RPP was compared to that of genetic algorithms, and, surprisingly, it emerged that genetic algorithms have some desirable properties: th expected fitness of the next generation bred from a selected population is independent of the population size; the expected fitness of the next generation is concentrated about the expected value; and there is a better bound on the improvement in expected fitness in one generation for a genetic algorithm than for the reinforcement learning algorithm. Hence the comparison of GAs with a reinforcement learning approach to a similiar problem revealed that GAs had some advantages.

26 Constructive Model Building

Chris Williams
Institute for Adaptive and Neural Computation
Division of Informatics, University of Edinburgh (GB)

Much work in statistical modelling consists of fitting the parameters of a given model to data, and can be carried out e.g. in a maximum likelihood or Bayesian setting. However, there is also the question of model structure choice, for example the number of components in a mixture model or a search over belief network structures.
In this talk I will give an overview of different methods that have been used in constructive model building approaches, where the structure of the model is built up depending on the data. I have identified three main approaches: (1) constructive learning by repeated re-representation, (2) constructive learning by data merging, (3) constructive learning as (greedy) search. Examples of models from these categories will be given.

27 SVM and VC Theory (Statistical Learning Theory)

Robert Williamson
Australian National University, Dept. of Telecommunications Engineering, Canberra (AU)

I presented a high-level view of the core insights of statistical learning theory. Considering as the goal of learning the minimization of expected risk, I considered three main induction principles for abstract learning algorithms: ERM (Empirical Risk Minimization), SRM (Structural Risk Minimization), DSRM (Data-Dependent SRM). I explained why convering numbers were the "right" qunatity to consider in analysing ERM. I explained the difficulties of DSRM and indicated how the so called luckiness framework was one way to rigorously reason about DSRM. I also pointed out relationships with the work of Jack Kieffer, Kylie Minogue and Britney Spears.