Report

on the Second Dagstuhl Seminar on

Theory and Practice of Machine Learning

January 6th - January 10th, 1997

Introduction

This Seminar brought together researchers from various branches of computer science whose work involves learning by machines. It provided an opportunity for theoreticans to measure their models and results against demands and data that arise in applied machine learning, and it provided an opportunity for researchers in applied machine learning to initiate a more systematic and rigorous theoretical investigation of new learning approaches that have emerged from experimental work on application problems.
One new learning paradigm that had received considerable attention at this Seminar is the Reinforcement Learning approach. This approach takes into account that in many practical machine learning scenarios no source of classified training examples is available to the learner. Another novel type of learning problems arises in the context of complex probabilistic models, which have turned out to be particularly useful for representing knowledge in a variety of real-world application domains. Apart from these two new directions the abstracts of this report demonstrate the vitality, sophistication and success of current work in theoretical and applied machine learning.
Most of the abstracts contain pointers to homepages where details of the reported work can be found. Pointers are also given to tutorials by Mike Jordan, Hans Simon und Rich Sutton, which provide introductions to some of the research areas discussed at this Seminar.
We would like to thank Lucas Palette for collecting these abstracts, and for providing pointers to the homepages of participants.

Wolfgang Maass

Participants

Peter Auer, Technische Universität Graz, Austria
Shai Ben-David, Technion, Haifa, Israel
Andreas Birkendorf, Universität Dortmund, Germany
Ivan Bratko, University of Ljubljana, Slovenia
Joachim M. Buhmann, Universität Bonn, Germany
Nicolò Cesa-Bianchi, Università degli Studi di Milano, Italy
Pawel Cichosz, Warsaw University of Technology, Poland
Eli Dichterman, Royal Holloway University of London, Great Britain
Thomas G. Dietterich, Oregon State University, USA
Carlos Domingo, Universidad Politécnica de Cataluña, Barcelona, Spain
Nicholas Dunkin, Royal Holloway University of London, Great Britain
Claude-Nicolas Fiechter, University of Pittsburgh, USA
Paul Fischer, Universität Paderborn, Germany
Ricard Gavaldà, Universidad Politécnica de Cataluña, Barcelona, Spain
Peter Grünwald, CWI Amsterdam, The Netherlands
Robert Holte, University of Ottawa, Canada
Michael I. Jordan, Massachusetts Institute of Technology, USA
Jyrki Kivinen, University of Helsinki, Finland
Norbert Klasner, Universität Dortmund, Germany
Eyal Kushilevitz, Technion, Haifa, Israel
Gabor Lugosi, Universidad - Pompeu Fabra, Barcelona, Spain
Klaus-Robert Müller, GMD First, Germany
Wolfgang Maass, Technische Universität Graz, Austria
Arnfried Ossen, Technische Universität Berlin, Germany
Lucas Paletta, Technische Universität Graz, Austria
Martin Riedmiller, Universität Karlsruhe, Germany
Dan Roth, Weizmann Institute, Israel
Stuart Russell, University of California at Berkeley, USA
Stefan Rüger, Technische Universität Berlin, Germany
Michael Schmitt, Technische Universität Graz, Austria
John Shawe-Taylor, Royal Holloway University of London, Great Britain
Hans Ulrich Simon, Universität Dortmund, Germany
Mark O. Stitson, Royal Holloway University of London, Great Britain
Richard S. Sutton, University of Massachusetts, USA
Naftali Tishby, The Hebrew University of Jerusalem, Israel

Bob Williamson, Australian National University, Canberra, Australia
Christopher M. Zach, Technische Universität Graz, Austria

Contents

PAUL FISCHER
Noise Models for PAC-Learning
PETER GRÜNWALD
MDL and the New Definition of Stochastic Complexity
ROBERT HOLTE
Application of Machine Learning to Oil Spill Detection
MICHAEL I. JORDAN
Graphical Models, Neural Networks and Variational Methods
NORBERT KLASNER
Learning Boolean Concepts Using "Minimum Equivalence Queries"
EYAL KUSHILEVITZ
On Learning Boolean Functions Represented as Automata
GABOR LUGOSI
Minimax Lower Bounds for the Two-Armed Bandit Problem
KLAUS-ROBERT MÜLLER
On-Line Learning in Changing Environments
WOLFGANG MAASS
How to Understand Learning in Biological Neural Systems and Pulsed VLSI?
ARNFRIED OSSEN
Weight Space Analysis and Forecast Uncertainty
LUCAS PALETTA
Temporal Difference Learning in Risk-Like Environments
MARTIN RIEDMILLER
Autonomously Learning Neural Controllers
DAN ROTH
Learning to Perform Knowledge-Intensive Inferences
STUART RUSSELL
Learning Complex Probabilistic Models

STEFAN RÜGER

Decimatable Boltzmann Machines: Efficient Inference and Learning

MICHAEL SCHMITT

The VC-Dimension of a Spiking Neuron

JOHN SHAWE-TAYLOR and NICHOLAS DUNKIN

Data Sensitive Analysis of Generalisation

JOHN SHAWE-TAYLOR and BOB WILLIAMSON

New Inductive Principles and "Luckiness"

HANS ULRICH SIMON

Tutorial on the PAC-Learning Model

HANS ULRICH SIMON

Presentation of an Open Problem: Analysis of On-Line Adversaries

MARKO O. STITSON

Support Vector Machines

RICHARD S. SUTTON

Reinforcement Learning Tutorial

RICHARD S. SUTTON

Exponentiated Gradient Methods for Reinforcement Learning

RICHARD S. SUTTON

TD Models: Modeling the World at a Mixture of Time Scales

NAFTALI TISHBY

Faithful Approximate Embeddings in Learning

BOB WILLIAMSON

Agnostic Learning of Nonconvex Classes of Functions

CHRISTOPHER M. ZACH

Markov Decision Problems with Complex Actions

Abstracts

On Learning from Multi-Instance Examples: Empirical Evaluation of a Theoretical Approach

by PETER AUER
http://www.cis.tu-graz.ac.at/igi/pauer/auer.html

We describe a practical algorithm for learning axis-parallel high-dimensional boxes from multi-instance examples. The first solution to this practical learning problem arising in drug design was given by Dietterich, Lathrop, and Lozano-Perez. A theoretical analysis was performed by Auer, Long, Srinivasan, and Tan. In this work we derive a competitive algorithm from theoretical considerations which is completely different from the approach taken by Dietterich et. al. Our algorithm uses for learning only simple statistics of the training data and avoids potentially hard computational problems which were solved by heuristics by Dietterich et. al. In empirical experiments our algorithm performs quite well although it does not reach the performance of the fine-tuned algorithm of Dietterich et. al. We conjecture that our approach can be fruitfully applied also to other learning problems where certain statistical assumptions are satisfied.

Combinatorial Variability of Vapnik-Chervonenkis Classes with Applications to Sample Compression Schemes

by SHAI BEN-DAVID (joint work with Ami Litman)

We define embeddings between concept classes that are meant to reflect certain aspects of their combinatorial structure. Furthermore, we introduce a notion of *universal concept classes* – classes into which any member of a given family of classes can be embedded. These universal classes play a role similar to that played in computational complexity by languages that are hard for a given complexity class. We show that classes of half-spaces in $\mathbb{R}^n$ are universal with respect to families of algebraically defined classes.

We present some combinatorial parameters along which the family of classes of a given VC-dimension can be grouped into sub-families. We use these

parameters to investigate the existence of embeddings and the scope of universality of classes. We view the formulation of these parameters and the related questions that they raise as a significant component in this work.

A second theme in our work is the notion of *Sample Compression Schemes*. Intuitively, a class C has a sample compression scheme if for any finite sample, labeled according to a member of C, there exists a short sub-sample so that the labels of the full sample can be reconstructed from this sub-sample.

By demonstrating the existence of certain compression schemes for the classes of half-spaces the existence of similar compression schemes for every class embeddable in half-spaces readily follows. We apply this approach to prove existence of compression schemes for all 'geometric concept classes'.

Using Computational Learning Strategies as a Tool for Combinatorial Optimization

by ANDREAS BIRKENDORF (joint work with Hans Ulrich Simon)
http://ls2-www.informatik.uni-dortmund.de/~birkendo/birkendo.html

We describe how a basic strategy from computational learning theory can be used to attack a class of NP-hard combinatorial optimization problems. It turns out that the learning strategy can be used as an iterative booster: given a solution to the combinatorial problem, we will start an efficient simulation of a learning algorithm which has a "good chance" to output an improved solution. This boosting technique is a new and surprisingly simple application of an existing learning strategy. It yields a novel heuristic approach to attack NP-hard optimization problems. It does not apply to each combinatorial problem, but we are able to exactly formalize some sufficient conditions. The new technique applies, for instance, to the problems of minimizing a deterministic finite automaton relative to a given domain, the analogous problem for ordered binary decision diagrams, and to graph coloring.

Acquisition and Transfer of Control Skill

by IVAN BRATKO
http://www-ai.ijs.si/ailab/bratko.html

Controlling a dynamic system, such as inverted pendulum, crane or aircraft, by a human operator, requires skill that the operator acquires through experience. In this talk, experiments with humans learning to control such systems are reviewed. In one particular experimental setting, the human is only shown the current state of the system (values of the system variables), but is not told what the actual controlled system is. In such a setting the human has no model of the system, not even a qualitative common sense model. Thus this is essentially learning to control a black box. One question for future research is to explore how well the techniques of reinforcement learning model the human skill acquisition process. Now assume that there is a skilled operator, then it is of interest to reconstruct the operator's subcognitive skill and transfer the skill into an automatic controller. This process, often called "behavioural cloning", is usually done by machine learning techniques where the control policy is learned as a function from system states to control actions. Results of experiments in behavioural cloning are summarised in this talk and the following open problems are identified:

- Although very successful controllers are often induced, the current approaches in general suffer from low yield of successful controllers. - Induced controllers are often brittle in the sense of being sensitive to small changes in the control task. - Induced controllers lack conceptual correspondence to the operator's plans, goals, causality relations and control loops. - How can skill from various operators be combined? - How can common sense causal models be exploited in skill reconstruction?

Review papers on this topic are:

I. Bratko, T. Urbancic, C. Sammut, Learning models of control skill: phenomena, results and problems. 13th IFAC World Congress, San Francisco, June-July, 1996

I. Bratko, T. Urbancic, C. Sammut, Behavioural cloning of control skill. In: Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications (eds. I.Bratko, M.Kubat, R.Michalski) Wiley 1997 (to appear)

Active Data Selection for Data Clustering

by JOACHIM M. BUHMANN (joint work with Thomas Hofmann)
http://www-dbv.informatik.uni-bonn.de/

Pairwise data clustering is an unsupervised learning problem where objects have to be grouped into clusters on the basis of their mutual dissimilarities. The clustering criterion favors groups with minimal average intra cluster dissimilarities. The $\mathcal{O}(N^2)$ growth of the number of dissimilarity values to characterize N objects suggests that an $\mathcal{O}(N)$ fraction of all dissimilarity values should be sufficient to reliably detect a cluster structure if it is present in the data. I describe a method to determine those dissimilarities which are most informative to find a good clustering solution. The data querying step and the data analysis procedure are tightly coupled to significantly speed up the discovery of data clusters.

Analysis of Two Gradient-Based Algorithms for On-Line Regression

by NICOLÒ CESA-BIANCHI
http://www.dsi.unimi.it/∼cesabian

In this paper we present a new analysis, in the on-line regression framework, of two algorithms: Gradient Descent and Exponentiated Gradient. Both these algorithms update their coefficients based on the gradient of the loss function; however, their update rules are substantially different. For each algorithm, we show general regression bounds for any convex loss function; furthermore, we show special bounds for the absolute and the square loss functions. This extends previous results by Kivinen and Warmuth. In the nonlinear regression case, we show general bounds for pairs of transfer and loss functions satisfying a certain condition. We apply this result to the Hellinger loss and the entropic loss in case of logistic regression (similar results, but only for the entropic loss, were also obtained by Helmbold et al. using a different analysis.)

On the Approximate Equivalence of Different Approaches to Reinforcement Learning Based on TD(λ)

by PAWEL CICHOSZ
http://www.ipe.pw.edu.pl/∼cichosz

This presentation reported some simple observations concerning the relationships between different computational techniques based on TD(λ). Specifically, two ideas described in the literature have been analyzed: *experience replay* and *replacing eligibility traces*.

Two closely related experience replay techniques used in other authors' prior work have been re-examined, referred to as *replayed TD* (RTD) and *backwards TD* (BTD). RTD(λ, n) is essentially an online learning method which performs at each time step a regular TD update, and then replays updates backwards for $n - 1$ previous states, so that a total of n updates are performed. In effect, each visited state is updated n times. BTD(λ) operates in offline mode, after the end of a trial updating backwards the predictions for all visited states. They can be both combined with general λ by using truncated TD(λ) returns for performing updates.

RTD(λ, n) and BTD(λ) have been shown to be approximately equivalent to standard TD($\lambda > 0$) with λ values set in a particular way. This is true even if RTD and BTD perform only TD(0) updates. This observations clarifies the effects of experience replay and shows that, in principle, it is no more powerful than regular TD(λ).

The definition and standard implementation of TD(λ) for $\lambda > 0$ is based on the well known idea of eligibility traces. For the classical, accumulating form of eligibility traces, each visit to a state adds to the current value of its trace, so that trace values depend on both the recency and frequency of visits. Recently replacing eligibility traces have been proposed, where visit frequency has no effect on trace values. Prior work on this novel form of eligibility traces has shown that it may be expected to perform better than accumulating traces, particularly for large λ. Whereas eligibility traces of any form represent a *backwards view* of TD(λ), the algorithm can be better understood using a *forwards view* to present its effects, based on TD(λ) returns. Such a forwards view for accumulating traces has been known for a number of years. Recently it has been also demonstrated that it may be of some practical usefulness as well, making it possible to implement TD(λ)

using an approximate, but more efficient technique than computationally demanding eligibility traces. In the work reported here the effects of replacing eligibility traces have been analyzed and a corresponding forwards view has been derived. The replacing analog of the TD(λ) return for time t has been identified as the m_t-step truncated TD(λ) return, with m_t set as the number of time steps between t and the next visit to the state visited at time t. The analysis provides an insightful illustration of the difference between the two forms of eligibility traces and suggests an efficient implementation of replacing TD(λ) without using traces.

More information about related research can be found via WWW at the following URL: *http://www.ipe.pw.edu.pl/∼cichosz* .

Learning with Partial Visibility

by ELI DICHTERMAN

The RFA (Restricted Focus of Attention) framework was introduced by Ben-David and Dichterman (COLT'93) to model partial visibility learning problems in which the learner interacts with the environment by focusing his attention on a specific set of features, then observing the values of these features for a randomly drawn example. In this model the learner can change his focus of attention during the learning process. A special case of the RFA setting is the k-RFA model, in which the learner is restricted to observing any set of k many attributes in each example. Under the uniform distribution over the Boolean cube $\{0,1\}^n$, a 1-RFA learner can compute good estimations of the Chow parameters (first order Fourier coefficients). Hence, since the Chow parameters uniquely define a perceptron (Bruck, 1990), the class of perceptrons is 1-RFA learnable under the uniform distribution. However, the following intriguing question remains open: Is it possible to efficiently obtain a good prediction rule from good approximations of the Chow parameters? In particular,is it possible to efficiently approximate a weight-based representation of a perceptron from good approximations of its Chow parameters? Notice that knowing the Chow parameters it is not hard to find a weak predictor. However, boosting is not always possible in the RFA setting (Birkendorf et al., COLT'96).

Reinforcement Learning for Job Shop Scheduling

by THOMAS G. DIETTERICH (joint work with W. Zhang)
http://www.cs.orst.edu/∼tgd
ftp://ftp.cs.orst.edu/pub/tgd/papers/ijcai95-jss.ps.gz

Job-shop scheduling can be formulated as a state-space search. The starting states are initial (critical path) schedules and the terminal states are feasible schedules. We applied the temporal difference learning algorithm TD(λ) to learn a value function for controlling this search. This can be done for a single instance of a scheduling problem by repeatedly solving the problem (with random exploration) and providing rewards proportional to the quality of the feasible solutions that are found. Our goal is to find the highest-quality feasible solution. Once the value function is learned, it can be used to solve this same problem instance without any search. However, this is not very useful, since by the time the value function is learned, we have found a good solution many times.

Our goal, however, is to learn a value function that can be applied to new problem instances. To achieve this, we first define a set of features that encode the state in a problem-instance-independent fashion. Next, we assume that our problem instances will be drawn from according to some source distribution. We draw a training sample from that distribution and learn to solve each of the problem instances in the training sample. We learn a value function (again by TD(λ)) that works well for all of these problem instances. We can then apply this value function to guide the solution of new problem instances.

Experiments with a NASA Space Shuttle scheduling problem show that our learned value function finds solutions of similar quality twice as fast as the best previous method for this problem. Alternatively, if given the same amount of time, it can find solutions that are significantly better than the best previous method.

Details can be found in our paper Zhang, W. and Dietterich, T. G. (1995), A Reinforcement Learning Approach to Job-Shop Scheduling,*Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal, Canada.

Hierarchical Reinforcement Learning

by THOMAS G. DIETTERICH

Reinforcement learning is very slow if the Markov decision problem (MDP) is very large. However, often the MDP has a hierarchical structure that can be exploited to make the solution tractable. For example, for a package delivery robot, the subproblem of navigating from one site to another can be solved independently of which packages the robot is currently carrying. In a flat search space, however, the value function would need to be learned separately for each possible combination of packages. Hierarchical value functions can address this problem. In this presentation, I reviewed work by Singh (1992), Kaelbling (1993), Lin (1993), and Dayan and Hinton (1993) and showed how each of these approaches constructed an overall value function from value functions for sub-Markov Decision Problems. I also described a proposed method, called Bureaucratic Reinforcement Learning (BRL), that shows promise of working better than previous methods.

In BRL, the value function is decomposed into a tree of functions. Each node in the tree corresponds to an action (the leaves are primitive actions and the root is a degenerate, most-abstract action). Each node also corresponds to a sub-goal (i.e., to a set of states where a given goal condition is true). Each node in the tree learns a Q function that represents the expected cumulative reward that would be received for performing the action defined by that node and then behaving optimally thereafter until the goal defined by the parent node is achieved. Each node computes its Q value by determining which of its children would be the most appropriate to execute. The Q value of this child is used to compute the Q value of the node (and this value is passed upward to the parent, recursively). The root node chooses to execute the child with the highest Q value; this child in turn executes its best child recursively until we reach a primitive leaf node at which a real action is performed. Then the entire action selection process is repeated in the new state.

Only the children of the root (i.e., the most abstract actions) estimate the global reward function. All other nodes in the tree have reward functions defined by their parent's goal. This permits these learned Q functions to be reused when the global task changes. However, it also introduces potential sub-optimalities into the resulting value function.

Hierarchical reinforcement learning introduces the following open questions.

(a) Are the proposed architectures coherent? (b) What conditions are required by a value function in order that it can be exactly hierarchically decomposed? (c) What conditions are required by a value function in order that it can be approximately hierarchically decomposed? (d) How hard is it to learn hierarchical value functions, and what learning algorithms would be most appropriate? (e) Can exploration policies exploit the hierarchy?

Partial Occam's Razor and its Applications

by CARLOS DOMINGO
(joint work with Tatsuie Tsukiji, and Osamu Watanabe)
http://www-lsi.upc.es/∼carlos

We introduce the notion of "partial Occam algorithm". A partial Occam algorithm produces a succinct hypothesis that is partially consistent with given examples, where the proportion of consistent examples is a bit more than half. By using this new notion, we propose one approach for obtaining a PAC learning algorithm. First, as shown in this paper, a partial Occam algorithm is equivalent to a weak PAC learning algorithm. Then by using boosting techniques of Schapire or Freund, we can obtain an ordinary PAC learning algorithm from this weak PAC learning algorithm. We demonstrate with some examples that some improvement is possible by this approach, in particuar in the hypotheses size.

First we obtain a PAC learning algorithm for k-DNF, which has similar sample complexity as Littlestone's Winnow, but produces hypothesis of size polynomial in d and $\log k$ for a k-DNF target with n variables and d terms (*Cf.* The hypothesis size of Winnow is $\mathcal{O}(n^k)$). Next we show that 1-decision lists of length d with n variables are (non-proper) PAC learnable by using $\mathcal{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}+16^d\log n(d+\log\log n)^2\right)$ examples within polynomial time w.r.t. n, 2^d, $1/\varepsilon$, and $\log 1/\delta$. Again, our sample size is the same as Winnow for the same problem but the hypothesis size is much smaller for some values of d.

We also show that both algorithms belong to the Statistical Query model of Kearns and therefore, are robust against classification and some attribute noise.

Expected Mistake Bound Model for On-Line Reinforcement Learning

by CLAUDE-NICOLAS FIECHTER
http://www.cs.pitt.edu/∼fiechter/

We propose a model of efficient on-line reinforcement learning based on the expected mistake bound framework introduced by Haussler, Littlestone and Warmuth. The measure of performance we use is the expected difference between the total reward received by the learning agent and that received by an agent behaving optimally from the start. We call this expected difference the *cumulative mistake* of the agent and we require that it "levels off" at a reasonably fast rate as the learning progresses. We show that this model is polynomially equivalent to the PAC model of off-line reinforcement learning that we introduced previously. In particular we show how an off-line PAC reinforcement learning algorithm can be transformed into an efficient on-line algorithm in a simple and practical way. An immediate consequence of this result is that the PAC algorithm for the general finite state-space reinforcement learning problem that we described in a previous work (COLT94) can be transformed into a polynomial on-line algorithm with guaranteed performances.

Noise Models for PAC-Learning

by PAUL FISCHER

The learning problems we consider can be briefly described as follows: One is given a domain X, a probability distribution D on X, and a set system $\mathcal{C} \subseteq 2^X$. The task is to find a *hypothesis* that approximates a *target* $C^\Lambda \in \mathcal{C}$. Information on C^Λ is provided by a *sample*, i.e., a sequence of pairs (x, l) where $x \in X$ is drawn under D and the label l indicates whether $x \in C^\Lambda$. In the *malicious noise* setting a certain fraction η, $0 < \eta < 1$, of the sample is replaced by maliciously chosen pairs. In the *classification noise* model only the label is inverted with probability η.

Our main objective is the sample size behaviour of learning algorithms in the malicious noise model. We are able to show that the minimum disagreement strategy (which is optimal in the classification noise model) is not sample size optimal in the malicious model if the final has to be deterministic. It is

outperformed by a randomized strategy computing a deterministic hypothesis.

We show that the use of randomised hypotheses improves the performance, especially if the noise rate η is high. It is shown that the information theoretic upper bound $\eta_{det} = \varepsilon/(1+\varepsilon)$ on the tolerable noise rate which holds for deterministic hypotheses can be replaced by $\eta_{rand} = 2\varepsilon/(1+2\varepsilon)$, where ε is the desired accuracy. Moreover we show that the sample size has a quadratic growth rate $1/\Delta^2$, where $\Delta = \eta_{rand} - \eta$, when the noise rate η approaches this bound.

We present an algorithm which uses randomised hypotheses for learning the powerset of d elements in the presence of malicious noise using an optimal sample size of order $d\varepsilon/\Delta^2$ (ignoring logarithmic factors) and tolerating a noise rate $\eta = \eta_{rand} - \Delta$. We complement this result by proving that this sample size is also necessary for any class $\mathcal{C}$. On the other hand, if the noise rate is considerably smaller than η_{rand} then one can learn with very small sample sizes. We present a master algorithm for this task.
The results have been obtained in collaboration with Nicolò Cesa-Bianchi, Eli Dichterman, Eli Shamir and Hans Ulrich Simon.

MDL and the New Definition of Stochastic Complexity

by PETER GRÜNWALD
http://www.cwi.nl/∼pdg

We give a short introduction of the new definition of Stochastic Complexity that Rissanen has recently (1996) proposed as a 'replacement' of the old one, which now is seen as being merely an approximation of the new one. The new definition of Stochastic Complexity sheds a clarifying light on the differences and similarities between Minimum Description Length and the Bayesian Approaches to learning, which has long been a topic of fierce discussions.

Application of Machine Learning to Oil Spill Detection

by ROBERT HOLTE
http://www.csi.uottawa.ca/~holte/

My talk centred on a real-world application of machine learning. The application is the detection of oil slicks in satellite radar images. The role of machine learning is to create a classifier that will decide if a region in an image should be brought to the attention of a human expert. After briefly describing the application, the talk described machine learning research issues that arose during the application which have not been adequately studied by the research community.

Graphical Models, Neural Networks and Variational Methods

by MICHAEL I. JORDAN
http://www.ai.mit.edu/projects/jordan.html

Graphical models provide a useful formalism for understanding neural networks and other network-based statistical and machine learning systems. The goal in all of these systems can be viewed as that of computing a probability distribution over hidden variables. For certain classes of architecture, this calculation can be carried out exactly and efficiently. For other architectures, the exact calculation is intractable and approximations must be developed. In this talk I describe the use of variational methods for the calculation of bounds of probabilities on graphs, exemplifying these methods with applications to Markovian probabilistic decision trees, factorial Hidden Markov models, Boltzmann machines and Bayesian belief networks with logistic or noisy-OR nodes.

The research is collaborative work with Zoubin Ghahramani (Toronto), Tommi Jaakkola (MIT) and Lawrence Saul (AT&T Research). You can find a compressed Postscript copy of the slides for my presentation at
ftp://psyche.mit.edu/pub/jordan/ai-stats-talk.ps.Z.

Learning Boolean Concepts Using "Minimum Equivalence Queries"

by NORBERT KLASNER
http://ls2-www.informatik.uni-dortmund.de/~klasner/

Query learning is done very often by using equivalence queries (EQ) or a so called minimal adequate teacher, which means using additionally membership queries ($MEMB$). The answer a learner gets to an equivalence query with a hypothesis h is an arbitrary couterexample x, if h is not equivalent to the concept class. This counterexample has to be considered as being chosen by an adversary rather than by a teacher. Since this environment seems to be too pessimistic, we restrict the set of possible counterexamples to counterexamples with a minimum Hamming weight. In the boolean case this are the counterexamples with a minimum number of ones. We call queries of this type "minimum equivalence queries" ($EQ^{\wedge}$). The use of $EQ^{\wedge}$ is motivated, for example, by different costs of counterexamples for the teacher. Also the teacher's intention may be to give valuable information, which holds in this setting for monotone functions.

The maximum length of an antichain which can be shattered, is a trivial lower bound for the worst case number of $EQ^{\wedge}$ queries needed to learn a concept class. We are searching for better (non-trivial) general lower bounds.

It is easy to learn the classes "Boolean treshold functions with 0-1-weights" and k-term-DNF in polynomial time using $EQ^{\wedge}$, which is known impossible using EQ only (unless P $\neq$ NP).

It is also not hard to learn k-quasi-Horn, which is a CNF with at most k unnegated literals in each clause, for constant k, using at most $\mathcal{O}(n^k m)$ $EQ^{\wedge}$ queries (where m is the number of clauses). But, for $k > 1$ this class is known to be as hard to learn as CNF using EQ and $MEMB$.

We are searching for boolean function classes, for which learning, using $EQ^{\wedge}$ is harder than using EQ and $MEMB$, and vice versa. We are also interested in structural results, e.g. similar to approximative finger prints.

On Learning Boolean Functions Represented as Automata

by EYAL KUSHILEVITZ
http://www.cs.technion.ac.il/∼eyalk

In this talk we survey some recent results in the *exact learning* model (i.e., learning with membership and equivalence queries). The common approach for the result that we survey is the following: we first prove that certain classes of functions that we are interested in can be represented using "small" automata (either deterministic finite automata, or so-called "multiplicity automata"). Then, we use algorithms that can learn these types of automata to learn these classes of functions.

Using this approach we show the learnability of polynomials over finite fields, classes of boxes in high dimension, generalized decision trees and other classes; hence, solving many of the open problems in this field.

Parts of this work were done jointly with A. Beimel, N. Bshouty, F. Bergadano, and S. Varricchio.

Minimax Lower Bounds for the Two-Armed Bandit Problem

by GABOR LUGOSI
http://www.econ.upf.es/∼lugosi

We obtain minimax lower bounds on the regret for the classical two-armed bandit problem. We offer a finite-sample minimax version of the well-known $\log n$ asymptotic lower bound of Lai and Robbins. Also, in contrast to the $\log n$ asymptotic results on the regret, we show that the minimax regret is achieved by mere random guessing under fairly mild conditions on the set of allowable configurations of the two arms. That is, we show that for *every* allocation rule and for *every* n, there is a configuration such that the regret at time n is at least $1 - \varepsilon$ times the regret of random guessing, where ε is any small positive constant.

On-Line Learning in Changing Environments

by KLAUS-ROBERT MÜLLER (joint work with Noboru Murata, Andreas Ziehe, and Shun-ichi Amari)
http://www.first.gmd.de/persons/Mueller.Klaus-Robert/pubs_klaus.html
http://neural-server.aston.ac.uk/nips95/workshop.html
http://www.first.gmd.de/persons/Mueller.Klaus-Robert.html

An adaptive on-line algorithm extending the learning of learning idea is proposed and theoretically motivated. Relying only on gradient flow information it can be applied to learning continuous functions or distributions, even when no explicit loss function is given and the Hessian is not available. Its efficiency is demonstrated for a non-stationary blind separation task of acoustic signals.

How to Understand Learning in Biological Neural Systems and Pulsed VLSI?

by WOLFGANG MAASS (joint work with Michael Schmitt)
http://www.cis.tu-graz.ac.at/igi/maass/

Recent results from neurobiology suggest that traditional models for neural information processing, such as McCulloch-Pitts neurons (i.e. threshold gates), multi-layer Perceptrons, sigmoidal gates etc., do not capture the essence of computation and learning in many biological neural networks. The main reason is that these models do not consider information to be encoded in temporal patterns of events, as it is the case in models for spiking neurons. (Similar mechanisms are explored with the technology of pulsed VLSI.)

This work is an attempt to capture essential features of new problems that arise if one analyzes computation and learning by a single neuron in temporal coding. In addition to the weights which model the plasticity of synaptic strength, the model includes variable transmission delays between neurons as programmable parameters.

Concerning the computational complexity of learning, the existence of a polynomial-time PAC-learning algorithm that uses an arbitrary polynomial-time computable hypothesis class is unlikely, because then the class DNF would be learnable by such an algorithm. We further show that for a spik-

ing neuron with delays from $\{0, 1\}$ PAC-learning with hypotheses from the same class is NP-hard even if all weights are kept fixed. This is particularly interesting because PAC-learning is feasible for a threshold gate, which can be considered as a spiking neuron where all delays have the same value.

These results show that temporal coding has a surprisingly large impact on the computational complexity of learning for single neurons.

Weight Space Analysis and Forecast Uncertainty

by ARNFRIED OSSEN (joint work with Stefan Rüger)
http://ini.cs.tu-berlin.de/∼ao

The usage of location information of weight vectors can help to overcome deficiencies of gradient based learning for neural networks. We study the non-trivial structure of weight space, i. e., symmetries of feedforward networks in terms of their corresponding groups. We find that these groups naturally act on and partition weight space into disjunct domains. We derive an algorithm to generate representative weight vectors in a fundamental domain. The analysis of the metric structure of the fundamental domain leads to a clustering method that exploits the natural metric of the fundamental domain. It can be implemented efficiently even for large networks. We used it to improve the assessment of forecast uncertainty for an already successful application of neural networks in the area of financial time series.

Temporal Difference Learning in Risk-Like Environments

by LUCAS PALETTA (joint work with F.J. Pineda)
http://www.icg.tu-graz.ac.at/cvgroup/staff/paletta.html
http://olympus.ece.jhu.edu/∼fernando/

Learning from temporal differences of a prediction function by TD(λ) enables autonomous agents to develop efficient policies in stochastic environments that deliver delayed reward. Encouraged by Tesauro's success with *TD-Gammon*, a reinforcement system is now defined to learn *Risk*-like games by self-play. In the *Simple Risk* environment, two agents compete for occupying all vertices of an arbitrarily connected graph by taking actions that cause changes in the local distribution of tokens on its nodes. An action can be ei-

ther an attack on an opponents vertex or a transfer of tokens between two of the agents own vertices. In the on-line setting, the TD(λ) estimator improves to evaluate the current board state, it asymptotically predicts the probability to win the game. The decision for the next move is a function of the predictor's evaluation of possible next states and transition probabilities that rule offensive conflicts. A special representation of the evaluation function results in weighting vertices according to their strategic importance due to topological relations. Success rates of trained versus random policies correlate with heuristic measures of graph sparsity by experimental evidence.

Autonomously Learning Neural Controllers

by MARTIN RIEDMILLER
http://www.ira.uka.de/∼riedml

Analytical controller design for dynamical systems can sometimes become arbitrarily difficult or even unfeasible, espescially due to nonlinearities in the dynamics or in the sensors and actuators. The goal of autonomously learning controllers is to overcome those problems by learning a tailored control strategy instead of designing one. The only training information thereby is given in terms of final success or failure, i.e. if the output target finally was reached or not. No external a priori knowledge about system behaviour or about a control policy is assumed.

The framework of dynamic programming offers a sound theoretical basis for the design of learning algorithms of a self learning controller. Our work focusses on applying and adapting these techniques for the use of a self learning, neural network based control system. The application to various nonlinear continuous control problems show the ability of the self-learning controller to learn high quality control strategies with a minimum of training information.

Learning to Perform Knowledge-Intensive Inferences

by DAN ROTH

We suggest an architecture for the development of systems that learn to perform knowledge-intensive inferences. Examples are language understanding related tasks, in which inference heavily depends on knowledge about the language and the world. The goal of this approach is to learn a knowledge base that can efficiently support various inferences. The main feature of this architecture is a scheme for enriching the primitive features the system received in its input. Features are enriched by incorporating additional knowledge, both syntactic – like potential part-of-speech information – and semantic – like synonyms, is-a relations and others. Complex features are generated by conjoining simple features that occur in close proximity. The enriched representation is the input to the learning algorithm. Learning is done using a Winnow-type algorithm, which can tolerate efficiently a large number of features, and is used also to discard irrelevant features. In this way the learning algorithm takes part also in the feature selection process. An important features of the algorithm is the use of Winnow to learn "thick" linear separator, which is shown to improve performance.

We experimented with this architecture on several large-scale inference problems in the language domain, and present results on context-sensitive spelling correction and text categorization. All inferences are supported by the same architecture, a learned knowledge base of size 10^6, and perform better that any known algorithms. In the context sensitive spelling domain we present also a preliminary learning-curve study and show that in many cases less than 50 examples are sufficient to support above 95

Learning Complex Probabilistic Models

by STUART RUSSELL
http://www/.cs.berkeley.edu/∼russell

Probabilistic (Bayesian) networks (PNs) provide a formal and compact representation for complex probability distributions. They are widely used in expert system applications, but until recently no methods were available for learning PNs from data, particularly in the case where some variables in the network are hidden. We derive a simple, gradient-based learning algorithm

for PNs, showing that the likelihood gradient with respect to the network parameters can be computed easily as a byproduct of loading the observed data onto the network. We then extend this approach to more complex models, including dynamic probabilistic networks (DPNs), which can be used to represent stochastic processes. DPNs appear to have significant advantages over HMMs in the representation of complex processes, because the number of parameters typically grows linearly rather than exponentially in the number of state variables.

Many open problems remain in this area, including 1) methods for fast, approximate inference to provide approximate gradient information; 2) finding more expressive representation classes, including hybrid discrete/continuous models, and deriving inference and learning algorithms for them; 3) analysing the sample and computational complexity of the PN learning problem; 4) using prior knowledge, such as qualitative probabilistic network annotations, to speed up the learning process; 5) finding effective heuristics for structure learning, particularly with hidden variables.

Decimatable Boltzmann Machines: Efficient Inference and Learning

by STEFAN RÜGER
http://ini.cs.tu-berlin.de/∼async

Activations in Boltzmann machines obey a distribution that is well-known from the canonical formalism of thermodynamics: the Boltzmann-Gibbs distribution. Marginalizing over hidden nodes, Boltzmann machines can be used to approximate probability distributions, store stochastic knowledge given by examples, and retrieve stored information. However, in the general case, the algorithms for these operations are NP-hard. We conclude that research should be directed away from the search for efficient inference and learning rules in general Boltzmann machines, and toward the design of special case algorithms.

Using the Fourier-Stieltjes transformation of the Boltzmann-Gibbs distribution, not only the key term

$$\langle s_a s_b \rangle = T \frac{\partial \log(Z_w)}{\partial w_{ab}}$$

of the learning rule but also the formula for inference can be expressed with the help of the partition sum Z_w.

Certain sufficiently interesting Boltzmann machines can be treated efficiently with the technique of decimation [Saul and Jordan 1994]. We have found a new decimation rule for binary Boltzmann machines and were able to show that there are no further decimation rules. This defines the set of *decimatable Boltzmann machines*. Exploiting the process of decimation allows us to calculate the partition sum of the system in an efficient way (linear in the number of nodes). Simple and efficient algorithms can be constructed as follows:

- Formulate the calculation of the log partition sum $\log(Z_w)$ as a feedforward network, where the weight vector w of the Boltzmann machine is the input vector.
- Do a forward and a backward pass in this network in order to calculate $\nabla \log(Z)$, thereby calculating all relevant terms $\langle s_a s_b \rangle = T(\nabla \log(Z))_{ab}$ together. This requires a slightly generalized version of the standard backpropagation algorithm.
- Using the formula for inference, the standard cost function (i.e., information gain) can be calculated. Being able to calculate both the cost function and its gradient allows us to apply any acceleration method for learning like quasi-Newton, conjugate gradient, learning rate adaptation, etc.

These algorithms yield exact results not only for learning but also for inference (as opposed to the traditional approach of Gibbs sampling). What is more, every decimatable Boltzmann machine can be converted to a network that calculates the partition sum. Thus decimatable Boltzmann machines have all benefits of recurrent and stochastic networks — and that using only deterministic feedforward networks!

The VC-Dimension of a Spiking Neuron

by MICHAEL SCHMITT (joint work with Wolfgang Maass)
http://www.cis.tu-graz.ac.at/igi/mschmitt/

Spiking neurons are models for the computational units in biological neural systems where information is considered to be encoded mainly in the temporal patterns of their activity. They provide a way of analyzing neural computation that is not captured by the traditional neuron models such as sigmoidal and threshold gates (or "Perceptrons").

The model we consider in this work has not only variable weights but also variable transmission delays between neurons as programmable parameters. For the input and output values two types of coding are taken into account: binary coding for the Boolean and analog coding for the real-valued domain.

We investigate the sample complexity of learning for a single spiking neuron within the framework of PAC-learnability. We show that the VC-dimension is $\Theta(n \log n)$ and, hence, strictly larger than that of a threshold gate which is $\Theta(n)$. In particular, the lower bound turns out to hold for binary coding and even holds if all weights are kept fixed. The upper bound is valid for the case of analog coding if weights *and* delays are programmable. Our upper bound also gives rise to an asymptotically tight bound for the number of Boolean functions that can be computed by a spiking neuron.

The results show that when considering temporal coding in single spiking neurons, variable delays lead to a much larger sample complexity than variable weights only.

Data Sensitive Analysis of Generalisation

by JOHN SHAWE-TAYLOR and NICHOLAS DUNKIN

The paper introduces some generalisations of Vapnik's method of structural risk minimisation (SRM). It considers the more general case when the hierarchy of classes is chosen in response to the data. A result is presented on the generalisation performance of classifiers with a "large margin". This theoretically explains the impressive generalisation performance of the maximal margin hyperplane algorithm of Vapnik and co-workers (which is the basis for their support vector machines). Results of experiments apply the same

principles to two layer neural networks were presented showing improved generalisation when a large margin to output weight ratio is demanded during training.

New Inductive Principles and "Luckiness"

by JOHN SHAWE-TAYLOR and BOB WILLIAMSON

We very briefly outlined the problem of the search for new inductive principles and posed some questions concerning a new framework introduced by the authors along with Peter Bartlett and Martin Anthony.

The vast majority of supervised learning algorithms in use are based on an inductive principle called empirical risk minimization. As Vapnik has pointed out, alternate principles exist (his maximum margin hyperplane and more generally the support vector machines are examples). An obvious problem is to find new ones. One does not only want a principle though; one would also like some performance bounds (on the generalisation error) in particular circumstances. In a paper in COLT96, the above mentioned authors developed a general framework called "luckiness" that allows one to achieve this goal in certain circumstances. An extended and revised version is available at *http://spigot.anu.edu.au/people/williams/lucky.ps* and as NeuroCOLT technical report NC-TR-96-053 at *ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports*. This framework was sketched and some examples given, including some minor variations of the maximum margin hyperplane.

The specific questions posed were concerned with the discovery of new principles within the luckiness framework, and whether indeed it is a sufficiently general framework. We do not believe it is the most general, but we do not have any examples of inductive principles for which the desired performance bounds on generalisation error are available, but for which it can be shown they can not be put into the luckiness framework (or perhaps that putting them into the framework necessarily results in far cruder performance bounds). Secondly, we were interested in whether particular heuristic strategies that researchers may have adopted to improve generalisation could be characterised by an appropriate luckiness function, hence making possible objective measures of its performance. The possibility of turning insights about the likely characteristics of a real learning problem into a measure

which could deliver hard guarantees of generalisation was proposed.

Tutorial on the PAC-Learning Model

by HANS ULRICH SIMON

The tutorial consisted of three parts. Part 1 introduced the basic pac-learning model, as proposed by Valiant in 1984, along with some extensions and variations of the model. Part 2 presented some central results in this model, namely the quantification of the information complexity of a learning problem in terms of the Vapnik-Chervonenkis dimension, the characterization of pac-learning algorithms as consistent hypotheses finders, and some cryptographical limitations on pac-learning. Part 3 described sample-size optimality as a challenge which often forces the design of new algorithmic ideas or new data structures. Available as a manuscript:
http://ls2-www.informatik.uni-dortmund.de/~klasner/dagstuhl/

Presentation of an Open Problem: Analysis of On-Line Adversaries

by HANS ULRICH SIMON

The analysis of adversaries, that corrupt a set of training examples by malicious noise in an on-line fashion, can be done in terms of a dynamic stochastic optimization problem, where the adversary pursues the long-term goal to fool a given learning strategy. A Markovian decision process describes how the noisy sample evolves. The actions of the adversary are the insertions of corrupted examples. It gets a reward of 1, if the noisy sample is finally converted by the learning strategy into a bad hypothesis.

This problem arises in the malicious pac-learning model. The methods of analysis are certainly related to methods that are heavily used in the field of reinforcement learning. If the best adversarial strategy cannot be derived analytically, there will be still be the chance to approximate it by means of reinforcement learning. This may give intuitive insights and help to analyze the problem.

Support Vector Machines

by MARK O. STITSON
http://wwwnew.cs.rhbnc.ac.uk/research/compint/areas/sv/

Support Vector Machines for classification tasks perform structural risk minimisation. They create a classifier with minimised VC Dimension. If the VC Dimension is low, the expected probability of error is low as well, which means good generalisation.

Support Vector Machines use a linear separating hyperplane to create a classifier, yet some problems cannot be linearly separated in the original input space. Support Vector Machines can non-linearly transform the original input space into a higher dimensional feature space. In this feature space it is trivial to find a linear optimal separating hyperplane. This hyperplane is optimal in the sense of being a maximal margin classifier with respect to the training data.

Open problems with this approach lie in two areas:

- The theoretical problem of which non-linear transformation to use.
- The practical problem of creating an efficient implementation as the basic algorithm has memory requirements which are squared with respect to the number of training examples and computational complexity which is also related to the number of training examples.

Reinforcement Learning Tutorial

by RICHARD S. SUTTON
http://envy.cs.umass.edu/People/sutton/sutton.html

Reinforcement learning is learning about, from, and while interacting with a environment in order to achieve a goal. In other words, it is a relatively direct model of the learning that people and animals do in their normal lives. In the last two decades, this age-old problem has come to be much better understood by integrating ideas from psychology, optimal control, artificial neural networks, and artificial intelligence. New methods and combinations of methods have enabled much better solutions to large-scale applications

than had been possible by all other means. This tutorial will provide a top-down introduction to the field, covering Markov decision processes and approximate value functions as the formulation of the problem, and dynamic programming, temporal-difference learning, and Monte Carlo methods as the principal solution methods. The role of neural networks, evolutionary methods, and planning will also be covered. The emphasis will be on understanding the capabilities and appropriate role of each of class of methods within in an integrated system for learning and decision making.

Exponentiated Gradient Methods for Reinforcement Learning TD Models

by RICHARD S. SUTTON (joint work with Doina Precup)
http://envy.cs.umass.edu/People/sutton/research-ideas.html#EG

We introduce and evaluate a natural extension of linear exponentiated gradient methods that makes them applicable to reinforcement learning problems. Just as these methods speed up supervised learning, we find that they can also increase the efficiency of reinforcement learning. Comparisons are made to conventional reinforcement learning methods on two test problems using CMAC function approximators and replace traces. On a small prediction task, EG methods showed no improvement, but on a larger control task (Mountain Car) they improved learning rate by approximately 25suggests that the difference may be due to the distribution of irrelevant features.

TD Models: Modeling the World at a Mixture of Time Scales

by RICHARD S. SUTTON (joint work with Doina Precup)
ftp://ftp.cs.umass.edu/pub/anw/pub/sutton/sutton-95.ps

Temporal-difference (TD) learning can be used not just to predict rewards, as is commonly done in reinforcement learning, but also to predict states, i.e., tolearn a model of the world's dynamics. We present theory and algorithms for intermixing TD models of the world at different levels of temporal abstraction within a single structure. Such multi-scale TD models can be used in model-based reinfrocement-learning architectures and dynamic programming methods in place of conventional Markov models. This enables

planning at higher and varied levels of abstraction, and, as such, may prove useful in formulating methods for hierarchical or multi-level planning and reinforcement learning. We treat primarily the prediction problem - that of learning a model and value function for the case of fixed agent behavior. Within this context, we establish the theoretical foundations of multi-scale models and derive TD algorithms for learning them. Two small computational experiments are presented to test and illustrate the theory. This work is an extension and generalization of the work of Singh (1992), Dayan (1993), and Sutton&Pinette (1985).

Faithful Approximate Embeddings in Learning

by NAFTALI TISHBY
http://www.cs.huji.ac.il/∼tishby

The issue of "good" representations of the input and hypothesis spaces in learning is a fundamental problem which has not been sufficiently addressed theoretically so far.

In this paper we consider the problem of embedding the input and hypotheses of boolean function classes in other classes, such that the natural metric structure of the two spaces is approximately preserved.

We first prove some general properties of such embeddings and then discuss possible approximate embedding in the class of "half-spaces" (single layer perceptrons) with dimension polynomial in the VC dimension of the original problem.

Our main result is that such an approximate embedding by half-spaces is possible for a class of problems, which we call "inforamtive", for which the dual problem (learning the input from labels of random hypotheses) has a similar VC-dimesion, and the variance of the generalization errors is bounded. We argue that many important learning problems are "informative" for typical distributions, e.g. geometric concepts, neural networks, decision trees, etc.

This is a rather surprising result, since there is a known combinatorial lower bound on the dimension of *exact* embedding by half-spaces, which is linear in the size of the space.

Agnostic Learning of Nonconvex Classes of Functions

by BOB WILLIAMSON
http://spigot.anu.edu.au/people/williams/home.html

It is known that the sample complexity of agnostic learning of convex classes of functions with squared loss is $O(1/\varepsilon)$ (ignoring log factors here, and elsewhere in this abstract). Furthermore, for nonconvex classes, a lower bound of $\Omega(1/\varepsilon^2)$ exists. In this talk I briefly outlined why a nonuniform bound of order $\mathcal{O}(1/\varepsilon)$ exists for nonconvex classes. The key property is how close the target conditional expectation is to a point of nonuniqueness of best approximation. I showed how some notions previously studied in the approximation theory literature can be used to determine bounds for certain classes. Some preliminary ideas on ways of stratifying the convex hull of a class of functions were also presented. This work was done in collaboration with Peter Bartlett, Jonathon Baxter and Phil Long.

Markov Decision Problems with Complex Actions

by CHRISTOPHER M. ZACH

I considered Markov decision problems (MDPs) where at every decision point the agent has to select a complex action, which is just a cartesian product of elemental actions. This situation can be interpreted in the following way: Instead of one agent making a complex decision use a team of agents where each member of the team selects an elemental action. The dynamics of the MDP and the immediate reward obviously depend on the whole vector of elemental actions. Some interesting questions are: (i) Are there any learning algorithms to train such teams of agents efficiently? (ii) What are the advantages/disadvantages of this approach? Some simple learning algorithms can be found, if some restricted form of communication is allowed between the agents.