Means, Correlations and Bounds
Martijn Leisink, Bert Kappen
The partition function for a Boltzmann machine can be bounded from above and below. We can use this to bound the means and the correlations. For networks with small weights, the values of these statistics can be restricted to non-trivial regions (i.e. a subset of [-1 , 1]). Experimental results show that reasonable bounding occurs for weight sizes where mean field expansions generally give good results.
****

Fragment Completion in Humans and Machines
David Jacobs, Bas Rokers, Archisman Rudra, Zili Liu
Partial information can trigger a complete memory. At the same time, human memory is not perfect. A cue can contain enough information to specify an item in memory, but fail to trigger that item. In the context of word memory, we present experiments that demonstrate some basic patterns in human memory errors. We use cues that consist of word frag- ments. We show that short and long cues are completed more accurately than medium length ones and study some of the factors that lead to this behavior. We then present a novel computational model that shows some of the ■exibility and patterns of errors that occur in human memory. This model iterates between bottom-up and top-down computations. These are tied together using a Markov model of words that allows memory to be accessed with a simple feature set, and enables a bottom-up process to compute a probability distribution of possible completions of word frag- ments, in a manner similar to models of visual perceptual completion.
****

Eye movements and the maturation of cortical orientation selectivity
Antonino Casile, Michele Rucci
Neural activity appears to be a crucial component for shaping the recep- tive ■elds of cortical simple cells into adjacent, oriented subregions alter- nately receiving ON- and OFF-center excitatory geniculate inputs. It is known that the orientation selective responses of V1 neurons are re■ned by visual experience. After eye opening, the spatiotemporal structure of neural activity in the early stages of the visual pathway depends both on the visual environment and on how the environment is scanned. We have used computational modeling to investigate how eye movements might affect the re■nement of the orientation tuning of simple cells in the pres- ence of a Hebbian scheme of synaptic plasticity. Levels of correlation be- tween the activity of simulated cells were examined while natural scenes were scanned so as to model sequences of saccades and ■xational eye movements, such as microsaccades, tremor and ocular drift. The speci■c patterns of activity required for a quantitatively accurate development of simple cell receptive ■elds with segregated ON and OFF subregions were observed during ■xational eye movements, but not in the presence of saccades or with static presentation of natural visual input. These re- sults suggest an important role for the eye movements occurring during visual ■xation in the re■nement of orientation selectivity.
****

Orientation-Selective aVLSI Spiking Neurons
Shih-Chii Liu, Jörg Kramer, Giacomo Indiveri, Tobi Delbrück, Rodney Douglas
We describe a programmable multi-chip VLSI neuronal system that can be used for exploring spike-based information processing models. The system consists of a silicon retina, a PIC microcontroller, and a transceiver chip whose integrate-and-■re neurons are connected in a soft winner-take-all architecture. The circuit on this multi-neuron chip ap- proximates a cortical microcircuit. The neurons can be con■gured for different computational properties by the virtual connections of a se- lected set of pixels on the silicon retina. The virtual wiring between the different chips is effected by an event-driven communication pro- tocol that uses asynchronous digital pulses, similar to spikes in a neu- ronal system. We used the multi-chip spike-based system to synthe- size orientation-tuned neurons using both a feedforward model and a feedback model. The performance of our analog hardware spiking model matched the experimental observations and digital simulations of continuous-valued neurons. The multi-chip VLSI system has advantages

over computer neuronal models in that it is real-time, and the computa- tional time does not scale with the size of the neuronal network.

****

On the Generalization Ability of On-Line Learning Algorithms

Nicolò Cesa-bianchi, Alex Conconi, Claudio Gentile

In this paper we show that on-line algorithms for classi■cation and re- gression can be naturally used to obtain hypotheses with good data- dependent tail bounds on their risk. Our results are proven without re- quiring complicated concentration-of-measure arguments and they hold for arbitrary on-line learning algorithms. Furthermore, when applied to concrete on-line algorithms, our results yield tail bounds that in many cases are comparable or better than the best known bounds.

****

Analysis of Sparse Bayesian Learning

Anita Faul, Michael Tipping

The recent introduction of the 'relevance vector machine' has effec(cid:173) tively demonstrated how sparsity may be obtained in generalised linear models within a Bayesian framework. Using a particular form of Gaussian parameter prior, 'learning' is the maximisation, with respect to hyperparameters, of the marginal likelihood of the data. This paper studies the properties of that objective func(cid:173) tion, and demonstrates that conditioned on an individual hyper(cid:173) parameter, the marginal likelihood has a unique maximum which is computable in closed form. It is further shown that if a derived 'sparsity criterion' is satisfied, this maximum is exactly equivalent to 'pruning' the corresponding parameter from the model.

****

A Hierarchical Model of Complex Cells in Visual Cortex for the Binocular Perception of Motion-in-Depth

Silvio Sabatini, Fabio Solari, Giulia Andreani, Chiara Bartolozzi, Giacomo Bisio

A cortical model for motion-in-depth selectivity of complex cells in the visual cortex is proposed. The model is based on a time ex(cid:173) tension of the phase-based techniques for disparity estimation. We consider the computation of the total temporal derivative of the time-varying disparity through the combination of the responses of disparity energy units. To take into account the physiological plau(cid:173) sibility, the model is based on the combinations of binocular cells characterized by different ocular dominance indices. The resulting cortical units of the model show a sharp selectivity for motion-in(cid:173) depth that has been compared with that reported in the literature for real cortical cells.

****

Motivated Reinforcement Learning

Peter Dayan

The standard reinforcement learning view of the involvement of neuromodulatory systems in instrumental conditioning in(cid:173) cludes a rather straightforward conception of motivation as prediction of sum future reward. Competition between actions is based on the motivating characteristics of their consequent states in this sense. Substantial, careful, experiments reviewed in Dickinson & Balleine, 12,13 into the neurobiology and psychol(cid:173) ogy of motivation shows that this view is incomplete. In many cases, animals are faced with the choice not between many dif(cid:173) ferent actions at a given state, but rather whether a single re(cid:173) sponse is worth executing at all. Evidence suggests that the motivational process underlying this choice has different psy(cid:173) chological and neural properties from that underlying action choice. We describe and model these motivational systems, and consider the way they interact.

****

Categorization by Learning and Combining Object Parts

Bernd Heisele, Thomas Serre, Massimiliano Pontil, Thomas Vetter, Tomaso Poggio

We describe an algorithm for automatically learning discriminative com- ponents

of objects with SVM classi■ers. It is based on growing image parts by minimizing theoretical bounds on the error probability of an SVM. Component-based face classi■ers are then combined in a second stage to yield a hierarchical SVM classi■er. Experimental results in face classi■cation show considerable robustness against rotations in depth and suggest performance at signi■cantly better level than other face detection systems. Novel aspects of our approach are: a) an algorithm to learn component-based classi■cation experts and their combination, b) the use of 3-D morphable models for training, and c) a maximum operation on the output of each component classi■er which may be relevant for bio- logical models of visual recognition.

****

Group Redundancy Measures Reveal Redundancy Reduction in the Auditory Pathway
Gal Chechik, Amir Globerson, M. Anderson, E. Young, Israel Nelken, Naftali Tishby

The way groups of auditory neurons interact to code acoustic in(cid:173) formation is investigated using an information theoretic approach. We develop measures of redundancy among groups of neurons, and apply them to the study of collaborative coding efficiency in two processing stations in the auditory pathway: the inferior colliculus (IC) and the primary auditory cortex (AI). Under two schemes for the coding of the acoustic content, acoustic segments coding and stimulus identity coding, we show differences both in information content and group redundancies between IC and AI neurons. These results provide for the first time a direct evidence for redundancy reduction along the ascending auditory pathway, as has been hy(cid:173) pothesized for theoretical considerations [Barlow 1959,2001]. The redundancy effects under the single-spikes coding scheme are signif(cid:173) icant only for groups larger than ten cells, and cannot be revealed with the redundancy measures that use only pairs of cells. The results suggest that the auditory system transforms low level rep(cid:173) resentations that contain redundancies due to the statistical struc(cid:173) ture of natural stimuli, into a representation in which cortical neu(cid:173) rons extract rare and independent component of complex acoustic signals, that are useful for auditory scene analysis.

****

Probabilistic Inference of Hand Motion from Neural Activity in Motor Cortex
Yun Gao, Michael Black, Elie Bienenstock, Shy Shoham, John Donoghue

Statistical learning and probabilistic inference techniques are used to in- fer the hand position of a subject from multi-electrode recordings of neu- ral activity in motor cortex. First, an array of electrodes provides train- ing data of neural ■ring conditioned on hand kinematics. We learn a non- parametric representation of this ■ring activity using a Bayesian model and rigorously compare it with previous models using cross-validation. Second, we infer a posterior probability distribution over hand motion conditioned on a sequence of neural test data using Bayesian inference. The learned ■ring models of multiple cells are used to de■ne a non- Gaussian likelihood term which is combined with a prior probability for the kinematics. A particle ■ltering method is used to represent, update, and propagate the posterior distribution over time. The approach is com- pared with traditional linear ■ltering methods; the results suggest that it may be appropriate for neural prosthetic applications.

****

The Noisy Euclidean Traveling Salesman Problem and Learning
Mikio Braun, Joachim Buhmann

We consider noisy Euclidean traveling salesman problems in the plane, which are random combinatorial problems with underlying structure. Gibbs sampling is used to compute average trajectories, which estimate the underlying structure common to all instances. This procedure requires identifying the exact relationship between permutations and tours. In a learning setting, the average trajec(cid:173) tory is used as a model to construct solutions to new instances sampled from the same source. Experimental results show that the average trajectory can in fact estimate the underlying structure and

that overfitting effects occur if the trajectory adapts too closely to a single instance.

****

Incorporating Invariances in Non-Linear Support Vector Machines

Olivier Chapelle, Bernhard Schölkopf

The choice of an SVM kernel corresponds to the choice of a rep(cid:173) resentation of the data in a feature space and, to improve per(cid:173) formance, it should therefore incorporate prior knowledge such as known transformation invariances. We propose a technique which extends earlier work and aims at incorporating invariances in non(cid:173) linear kernels. We show on a digit recognition task that the pro(cid:173) posed approach is superior to the Virtual Support Vector method, which previously had been the method of choice.

****

Sampling Techniques for Kernel Methods

Dimitris Achlioptas, Frank Mcsherry, Bernhard Schölkopf

We propose randomized techniques for speeding up Kernel Principal Component Analysis on three levels: sampling and quantization of the Gram matrix in training, randomized rounding in evaluating the kernel expansions, and random projections in evaluating the kernel itself. In all three cases, we give sharp bounds on the accuracy of the obtained ap- proximations. Rather intriguingly, all three techniques can be viewed as instantiations of the following idea: replace the kernel function by a “randomized kernel” which behaves like

****

Reinforcement Learning and Time Perception -- a Model of Animal Experiments

Jonathan Shapiro, J. Wearden

Animal data on delayed-reward conditioning experiments shows a striking property - the data for different time intervals collapses into a single curve when the data is scaled by the time interval. This is called the scalar property of interval timing. Here a simple model of a neural clock is presented and shown to give rise to the scalar property. The model is an accumulator consisting of noisy, linear spiking neurons. It is analytically tractable and contains only three parameters. When coupled with reinforcement learning it simulates peak procedure experiments, producing both the scalar property and the pattern of single trial covariances.

****

Hyperbolic Self-Organizing Maps for Semantic Navigation

Jorg Ontrup, Helge Ritter

We introduce a new type of Self-Organizing Map (SOM) to navigate in the Semantic Space of large text collections. We propose a “hyper- bolic SOM” (HSOM) based on a regular tesselation of the hyperbolic plane, which is a non-euclidean space characterized by constant negative gaussian curvature. The exponentially increasing size of a neighborhood around a point in hyperbolic space provides more freedom to map the complex information space arising from language into spatial relations. We describe experiments, showing that the HSOM can successfully be applied to text categorization tasks and yields results comparable to other state-of-the-art methods.

****

Probabilistic Abstraction Hierarchies

Eran Segal, Daphne Koller, Dirk Ormoneit

Many domains are naturally organized in an abstraction hierarchy or taxonomy, where the instances in “nearby” classes in the taxonomy are similar. In this pa- per, we provide a general probabilistic framework for clustering data into a set of classes organized as a taxonomy, where each class is associated with a prob- abilistic model from which the data was generated. The clustering algorithm simultaneously optimizes three things: the assignment of data instances to clus- ters, the models associated with the clusters, and the structure of the abstraction hierarchy. A unique feature of our approach is that it utilizes global optimization algorithms for both of the last two steps, reducing the sensitivity to noise and the propensity to local maxima that are characteristic of algorithms such

as hierarchi- cal agglomerative clustering that only take local steps. We provid e a theoretical analysis for our algorithm, showing that it converges to a local maximum of the joint likelihood of model and data. We present experimental resu lts on synthetic data, and on real data in the domains of gene expression and te xt.

****

Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade
Paul Viola, Michael Jones

This paper develops a new approach for extremely fast detection in do- mains whe re the distribution of positive and negative examples is highly skewed (e.g. fac e detection or database retrieval). In such domains a cascade of simple classi■e rs each trained to achieve high detection rates and modest false positive rates can yield a ■nal detector with many desir- able features: including high detecti on rates, very low false positive rates, and fast performance. Achieving extreme ly high detection rates, rather than low error, is not a task typically addresse d by machine learning al- gorithms. We propose a new variant of AdaBoost as a me chanism for training the simple classi■ers used in the cascade. Experimental res ults in the domain of face detection show the training algorithm yields sig- ni■ cant improvements in performance over conventional AdaBoost. The ■nal face detec tion system can process 15 frames per second, achieves over 90% detection, and a false positive rate of 1 in a 1,000,000.

****

Agglomerative Multivariate Information Bottleneck
Noam Slonim, Nir Friedman, Naftali Tishby

The information bottleneck method is an unsupervised model independent data organization technique. Given a joint distribution peA, B), this method con(cid:173) structs a new variable T that extracts partitions, or clusters, o ver the values of A that are informative about B. In a recent paper, we intr oduced a general princi(cid:173) pled framework for multivariate extensions of t he information bottleneck method that allows us to consider multiple systems of data partitions that are inter-related. In this paper, we present a new family of simple agglomerative algorithms to construct such systems of inte r-related clusters. We analyze the behavior of these algorithms and apply them to several real-life datasets.

****

K-Local Hyperplane and Convex Distance Nearest Neighbor Algorithms
Pascal Vincent, Yoshua Bengio

Guided by an initial idea of building a complex (non linear) decision surface wi th maximal local margin in input space, we give a possible geometrical intuition as to why K-Nearest Neighbor (KNN) algorithms often perform more poorly than SV Ms on classi■cation tasks. We then propose modi■ed K-Nearest Neighbor algorithms to overcome the per- ceived problem. The approach is similar in spirit to Tange nt Distance, but with invariances inferred from the local neighborhood rather th an prior knowledge. Experimental results on real world classi■cation tasks sug- gest that the modi■ed KNN algorithms often give a dramatic improve- ment over st andard KNN and perform as well or better than SVMs.

****

Using Vocabulary Knowledge in Bayesian Multinomial Estimation
Thomas Griffiths, Joshua Tenenbaum

Estimating the parameters of sparse multinomial distributions is an important co mponent of many statistical learning tasks. Recent approaches have used uncertai nty over the vocabulary of symbols in a multinomial distribution as a means of a ccounting for sparsity. We present a Bayesian approach that allows weak prior kn owledge, in the form of a small set of approximate candidate vocabularies, to be used to dramatically improve the resulting estimates. We demonstrate these impr ovements in applications to text compres(cid:173) sion and estimating distributi ons over words in newsgroup data.

****

Bayesian time series classification
Peter Sykacek, Stephen J. Roberts

This paper proposes an approach to classi■cation of adjacent segments of a time series as being either of classes. We use a hierarchical model that consists of a feature extraction stage and a generative classi■er which is built on top of these features. Such two stage approaches are often used in signal and image pro cessing. The novel part of our work is that we link these stages probabilistical ly by using a latent feature space. To use one joint model is a Bayesian require ment, which has the advantage to fuse information according to its certainty. Th e classi■er is implemented as hidden Markov model with Gaussian and Multinomial observation distributions de■ned on a suitably chosen representation of autoregr essive models. The Markov dependency is mo- tivated by the assumption that succe ssive classi■cations will be corre- lated. Inference is done with Markov chain M onte Carlo (MCMC) tech- niques. We apply the proposed approach to synthetic data and to classi- ■cation of EEG that was recorded while the subjects performed di fferent cognitive tasks. All experiments show that using a latent feature space results in a signi■cant improvement in generalization accuracy. Hence we expect that this idea generalizes well to other hierarchical models.
****

Computing Time Lower Bounds for Recurrent Sigmoidal Neural Networks
M. Schmitt
Recurrent neural networks of analog units are computers for real(cid:173) valu ed functions. We study the time complexity of real computa(cid:173) tion in general recurrent neural networks. These have sigmoidal, linear, and pr oduct units of unlimited order as nodes and no re(cid:173) strictions on the weights. For networks operating in discrete time, we exhibit a family of functions with arbitrarily high complexity, and we derive almost tight b ounds on the time required to compute these functions. Thus, evidence is give n of the computational lim(cid:173) itations that time-bounded analog recurr ent neural networks are subject to.
****

Playing is believing: The role of beliefs in multi-agent learning
Yu-Han Chang, Leslie Pack Kaelbling
We propose a new classi■cation for multi-agent learning algorithms, with each le ague of players characterized by both their possible strategies and possible bel iefs. Using this classi■cation, we review the optimality of ex- isting algorithm s, including the case of interleague play. We propose an incremental improvement to the existing algorithms that seems to achieve average payoffs that are at le ast the Nash equilibrium payoffs in the long- run against fair opponents.
****

Probabilistic principles in unsupervised learning of visual structure: human dat a and a model
Shimon Edelman, Benjamin Hiles, Hwajin Yang, Nathan Intrator
To ■nd out how the representations of structured visual objects depend on the co -occurrence statistics of their constituents, we exposed subjects to a set of co mposite images with tight control exerted over (1) the condi- tional probabiliti es of the constituent fragments, and (2) the value of Bar- low’s criterion of “s uspicious coincidence” (the ratio of joint probability to the product of margina ls). We then compared the part veri■cation re- sponse times for various probe/ta rget combinations before and after the exposure. For composite probes, the speed up was much larger for tar- gets that contained pairs of fragments perfectly pre dictive of each other, compared to those that did not. This effect was modulated by the sig- ni■cance of their co-occurrence as estimated by Barlow’s criterion. For lone-fragment probes, the speedup in all conditions was generally lower tha n for composites. These results shed light on the brain’s strategies for unsuper vised acquisition of structural information in vision.
****

Predictive Representations of State
Michael Littman, Richard S. Sutton
We show that states of a dynamical system can be usefully repre(cid:173) sented by multi-step, action-conditional predictions of future ob(cid:173) servations. State representations that are grounded in data in this way may be easier to lea

rn, generalize better, and be less depen(cid:173) dent on accurate prior models than, for example, POMDP state representations. Building on prior work by Jaeger and by Rivest and Schapire, in this paper we compare and contrast a linear spe(cid:173) cialization of the predictive approach with the state representa(cid:173) tions used in POMDPs and in k-order Markov models. Ours is the first specific formulation of the predictive idea that includes both stochasticity and actions (controls). We show that any system has a linear predictive state representation with number of predictions no greater than the number of states in its minimal POMDP model.

****

Discriminative Direction for Kernel Classifiers

Polina Golland

In many scientific and engineering applications, detecting and under- standing differences between two groups of examples can be reduced to a classical problem of training a classifier for labeling new examples while making as few mistakes as possible. In the traditional classifi- cation setting, the resulting classifier is rarely analyzed in terms of the properties of the input data captured by the discriminative model. How- ever, such analysis is crucial if we want to understand and visualize the detected differences. We propose an approach to interpretation of the sta- tistical model in the original feature space that allows us to argue about the model in terms of the relevant changes to the input vectors. For each point in the input space, we define a discriminative direction to be the direction that moves the point towards the other class while introducing as little irrelevant change as possible with respect to the classifier func- tion. We derive the discriminative direction for kernel-based classifiers, demonstrate the technique on several examples and briefly discuss its use in the statistical shape analysis, an application that originally motivated this work. 1 Introduction

****

Contextual Modulation of Target Saliency

Antonio Torralba

The most popular algorithms for object detection require the use of exhaustive spatial and scale search procedures. In such approaches, an object is defined by means of local features. fu this paper we show that including contextual information in object detection pro(cid:173) cedures provides an efficient way of cutting down the need for exhaustive search. We present results with real images showing that the proposed scheme is able to accurately predict likely object classes, locations and sizes.

****

On Kernel-Target Alignment

Nello Cristianini, John Shawe-Taylor, André Elisseeff, Jaz Kandola

We introduce the notion of kernel-alignment, a measure of similar(cid:173) ity between two kernel functions or between a kernel and a target function. This quantity captures the degree of agreement between a kernel and a given learning task, and has very natural interpre(cid:173) tations in machine learning, leading also to simple algorithms for model selection and learning. We analyse its theoretical properties, proving that it is sharply concentrated around its expected value, and we discuss its relation with other standard measures of per(cid:173) formance. Finally we describe some of the algorithms that can be obtained within this framework, giving experimental results show(cid:173) ing that adapting the kernel to improve alignment on the labelled data significantly increases the alignment on the test set, giving improved classification accuracy. Hence, the approach provides a principled method of performing transduction.

****

Escaping the Convex Hull with Extrapolated Vector Machines

Patrick Haffner

Maximum margin classifiers such as Support Vector Machines (SVMs) critically depends upon the convex hulls of the training samples of each class, as they implicitly search for the minimum distance between the conv

ex hulls. We propose Extrapolated Vec(cid:173) tor Machines (XVMs) which r ely on extrapolations outside these convex hulls. XVMs improve SVM general ization very significantly on the MNIST [7] OCR data. They share simila rities with the Fisher discriminant: maximize the inter-class margin wh ile mini(cid:173) mizing the intra-class disparity.

****

Algorithmic Luckiness

Ralf Herbrich, Robert C. Williamson

In contrast to standard statistical learning theory which studies unifo rm bounds on the expected error we present a framework that exploits the spec ific learning algorithm used. Motivated by the luckiness framework [8] 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 hypothe(cid:173) ses in a given hypothesis space it is only necessary to cover the functions which could have been learne d using the fixed learning algorithm. We show how the resulting frame work relates to the VC, luckiness and compression frameworks. Finally, we present an application of this framework to the maximum margin algorith m for linear classifiers which results in a bound that exploits both the mar gin and the distribution of the data in feature space.

****

Optimising Synchronisation Times for Mobile Devices

Neil Lawrence, Antony I. T. Rowstron, Christopher Bishop, Michael J. Taylor

With the increasing number of users of mobile computing devices (e.g. personal digital assistants) and the advent of third generation mobile phones, wireless communications are becoming increasingly important. Many applications rely on the device maintaining a replica of a data-structure which is store d on a server, for exam(cid:173) ple news databases, calendars and e-mail. ill this paper we explore the question of the optimal strategy for synchronisin g such replicas. We utilise probabilistic models to represent how the data-stru ctures evolve and to model user behaviour. We then formulate objective fun ctions which can be minimised with respect to the synchronisa(cid:173) tion timi ngs. We demonstrate, using two real world data-sets, that a user can obtain mo re up-to-date information using our approach.

****

Spike timing and the coding of naturalistic sounds in a central auditory area of songbirds

B. Wright, Kamal Sen, William Bialek, A. Doupe

In nature, animals encounter high dimensional sensory stimuli that have complex statistical and dynamical structure. Attempts to study the neu- ral coding of th ese natural signals face challenges both in the selection of the signal ensemble and in the analysis of the resulting neural responses. For zebra ■nches, natura listic stimuli can be de■ned as sounds that they encounter in a colony of conspe ci■c birds. We assembled an ensemble of these sounds by recording groups of 10-4 0 zebra ■nches, and then ana- lyzed the response of single neurons in the songbi rd central auditory area (■eld L) to continuous playback of long segments from t his ensemble. Following methods developed in the ■y visual system, we measured t he information that spike trains provide about the acoustic stimulus with- out a ny assumptions about which features of the stimulus are relevant. Preliminary re sults indicate that large amounts of information are carried by spike timing, wi th roughly half of the information accessible only at time resolutions better th an 10 ms; additional information is still be- ing revealed as time resolution is improved to 2 ms. Information can be decomposed into that carried by the lockin g of individual spikes to the stimulus (or modulations of spike rate) vs. that c arried by timing in spike patterns. Initial results show that in ■eld L, tempora l patterns give at least

****

A Variational Approach to Learning Curves

Dörthe Malzahn, Manfred Opper

We combine the replica approach from statistical physics with a varia- tional ap

proach to analyze learning curves analytically. We apply the method to Gaussian process regression. As a main result we derive ap- proximative relations between empirical error measures, the generaliza- tion error and the posterior variance .

****

Why Neuronal Dynamics Should Control Synaptic Learning Rules

Jesper Tegnér, Ádám Kepecs

Hebbian learning rules are generally formulated as static rules. Un(cid:173)der changing condition (e.g. neuromodulation, input statistics) most rules are sen sitive to parameters. In particular, recent work has focused on two different f ormulations of spike-timing-dependent plasticity rules. Additive STDP [1] is re markably versatile but also very fragile, whereas multiplicative STDP [2, 3] is more ro(cid:173)bust but lacks attractive features such as synaptic competitio n and rate stabilization. Here we address the problem of robustness in the add itive STDP rule. We derive an adaptive control scheme, where the learning funct ion is under fast dynamic control by post(cid:173)synaptic activity to stabiliz e learning under a variety of conditions. Such a control scheme can be implemen ted using known biophysical mechanisms of synapses. We show that this adaptive rule makes the addit ive STDP more robust. Finally, we give an example how met a plasticity of the adaptive rule can be used to guide STDP into different type of learning regimes.

****

Latent Dirichlet Allocation

David Blei, Andrew Ng, Michael Jordan

We propose a generative model for text and other collections of dis(cid:173)cre te data that generalizes or improves on several previous models including naive Bayes/unigram, mixture of unigrams [6], and Hof(cid:173)mann's aspect mode l, also known as probabilistic latent semantic indexing (pLSI) [3]. In the context of text modeling, our model posits that each document is generated as a mixture of topics, where the continuous-valued mixture proportions are distributed as a latent Dirichlet random variable. Inf erence and learning are carried out efficiently via variational algorithms . We present em(cid:173)pirical results on applications of this model to problems in text modeling, collaborative filtering, and text classifica tion.

****

A Bayesian Network for Real-Time Musical Accompaniment

Christopher Raphael

We describe a computer system that provides a real-time musi(cid:173)cal accompaniment for a live soloist in a piece of non-improvised music for soloist and accompaniment. A Bayesian network is devel(cid:173)oped that represents the joint distribution on the times at which the solo and accompaniment notes are played, relating the two parts through a layer of hidden variables. The network is first con(cid:173)structed using the rh ythmic information contained in the musical score. The network is then train ed to capture the musical interpre(cid:173)tations of the soloist and accompani st in an off-line rehearsal phase. During live accompaniment the learned distri bution of the network is combined with a real-time analysis of the solois t's acoustic sig(cid:173)nal, performed with a hidden Markov model, to generate a musi(cid:173)cally principled accompaniment that respects all av ailable sources of knowledge. A live demonstration will be provided.

****

Efficiency versus Convergence of Boolean Kernels for On-Line Learning Algorithms

Roni Khardon, Dan Roth, Rocco A. Servedio

We study online learning in Boolean domains using kernels which cap- ture featur e expansions equivalent to using conjunctions over basic fea- tures. We demonstr ate a tradeoff between the computational ef■ciency with which these kernels can be computed and the generalization abil- ity of the resulting classi■er. We ■rst describe several kernel functions which capture either limited forms of conjunc tions or all conjunctions. We show that these kernels can be used to ef■ciently

run the Percep- tron algorithm over an exponential number of conjunctions; howev er we also prove that using such kernels the Perceptron algorithm can make an ex ponential number of mistakes even when learning simple func- tions. We also cons ider an analogous use of kernel functions to run the multiplicative-update Winno w algorithm over an expanded feature space of exponentially many conjunctions. W hile known upper bounds imply that Winnow can learn DNF formulae with a polynomi al mistake bound in this setting, we prove that it is computationally hard to si mulate Win- now's behavior for learning DNF over such a feature set, and thus th at such kernel functions for Winnow are not ef■ciently computable.
****

Prodding the ROC Curve: Constrained Optimization of Classifier Performance
Michael C. Mozer, Robert Dodier, Michael Colagrosso, Cesar Guerra-Salcedo, Richa rd Wolniewicz
When designing a two-alternative classi■er, one ordinarily aims to maximize the classi■er's ability to discriminate between members of the two classes. We descr ibe a situation in a real-world business application of machine-learning predict ion in which an additional constraint is placed on the nature of the solu- tion: that the classi■er achieve a speci■ed correct acceptance or correct rejection r ate (i.e., that it achieve a ■xed accuracy on members of one class or the other) . Our domain is predicting churn in the telecommunications industry. Churn refer s to customers who switch from one service provider to another. We pro- pose fou r algorithms for training a classi■er subject to this domain constraint, and pre sent results showing that each algorithm yields a reliable improvement in perfor mance. Although the improvement is modest in magnitude, it is nonethe- less impr essive given the dif■culty of the problem and the ■nancial return that it achiev es to the service provider.
****

Natural Language Grammar Induction Using a Constituent-Context Model
Dan Klein, Christopher D. Manning
This paper presents a novel approach to the unsupervised learning of syn- tactic analyses of natural language text. Most previous work has focused on maximizing likelihood according to generative PCFG models. In con- trast, we employ a simp ler probabilistic model over trees based directly on constituent identity and li near context, and use an EM-like iterative procedure to induce structure. This m ethod produces much higher qual- ity analyses, giving the best published results on the ATIS dataset. 1 Overview
****

Estimating Car Insurance Premia: a Case Study in High-Dimensional Data Inference
Nicolas Chapados, Yoshua Bengio, Pascal Vincent, Joumana Ghosn, Charles Dugas, I chiro Takeuchi, Linyan Meng
Estimating insurance premia from data is a difficult regression problem for seve ral reasons: the large number of variables, many of which are .discrete, and the very peculiar shape of the noise distri(cid:173) bution, asymmetric with fat ta ils, with a large majority zeros and a few unreliable and very large values. We compare several machine learning methods for estimating insurance premia, and te st them on a large data base of car insurance policies. We find that func(cid:173) tion approximation methods that do not optimize a squared loss, like Support Vector Machines regression, do not work well in this context. Compared methods i nclude decision trees and generalized linear models. The best results are obtain ed with a mixture of experts, which better identifies the least and most risky c ontracts, and allows to reduce the median premium by charging more to the most r isky customers.
****

Classifying Single Trial EEG: Towards Brain Computer Interfacing
Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller
Driven by the progress in the ■eld of single-trial analysis of EEG, there is a g rowing interest in brain computer interfaces (BCIs), i.e., systems that enable h uman subjects to control a computer only by means of their brain signals. In a p seudo-online simulation our BCI detects upcoming ■nger movements in a natural ke yboard typing condition and predicts their lat- erality. This can be done on ave

rage 100–230 ms before the respective key is actually pressed, i.e., long before the onset of EMG. Our approach is appealing for its short response time and high classi■cation accuracy (>96%) in a binary decision where no human training is involved. We compare discriminative classi■ers like Support Vector Machines (SVMs) and different variants of Fisher Discriminant that possess favorable reg- ularization properties for dealing with high noise cases (inter-trial vari- ability).

****

Kernel Logistic Regression and the Import Vector Machine

Ji Zhu, Trevor Hastie

The support vector machine (SVM) is known for its good performance in binary classi■cation, but its extension to multi-class classi■cation is still an on-going research issue. In this paper, we propose a new approach for classi■cation, called the import vector machine (IVM), which is built on kernel logistic regression (KLR). We show that the IVM not only per- forms as well as the SVM in binary classi■cation, but also can naturally be generalized to the multi-class case. Furthermore, the IVM provides an estimate of the underlying probability. Similar to the “support points” of the SVM, the IVM model uses only a fraction of the training data to index kernel basis functions, typically a much smaller fraction than the SVM. This gives the IVM a computational advantage over the SVM, especially when the size of the training data set is large.

****

A Model of the Phonological Loop: Generalization and Binding

Randall O'Reilly, R. Soto

We present a neural network model that shows how the prefrontal cortex, interacting with the basal ganglia, can maintain a sequence of phonological information in activation-based working memory (i.e., the phonological loop). The primary function of this phono(cid:173) logical loop may be to transiently encode arbitrary bindings of information necessary for tasks - the combinatorial expressive power of language enables very flexible binding of essentially ar(cid:173) bitrary pieces of information. Our model takes advantage of the closed-class nature of phonemes, which allows different neural rep(cid:173) resentations of all possible phonemes at each sequential position to be encoded. To make this work, we suggest that the basal ganglia provide a region-specific update signal that allocates phonemes to the appropriate sequential coding slot. To demonstrate that flexi(cid:173) ble, arbitrary binding of novel sequences can be supported by this mechanism, we show that the model can generalize to novel se(cid:173) quences after moderate amounts of training.

****

Thin Junction Trees

Francis Bach, Michael Jordan

We present an algorithm that induces a class of models with thin junction trees—models that are characterized by an upper bound on the size of the maximal cliques of their triangulated graph. By ensuring that the junction tree is thin, inference in our models remains tractable throughout the learning process. This allows both an ef■cient implementation of an iterative scaling parameter estimation algorithm and also ensures that inference can be performed ef■ciently with the ■nal model. We illustrate the approach with applications in handwritten digit recognition and DNA splice site detection.

****

Linking Motor Learning to Function Approximation: Learning in an Unlearnable Force Field

O. Donchin, Reza Shadmehr

Reaching movements require the brain to generate motor com- mands that rely on an internal model of the task's dynamics. Here we consider the errors that subjects make early in their reaching trajectories to various targets as they learn an internal model. Us- ing a framework from function approximation, we argue that the sequence of errors should reflect the process of gradient descent. If so, then the sequence of errors should obey hidden state transitions of a simple dynam

ical system. Fitting the system to human data, we find a surprisingly good fit a ccounting for 98% of the variance. This allows us to draw tentative conclusions about the basis ele- ments used by the brain in transforming sensory space to mo tor commands. To test the robustness of the results, we estimate the shape of th e basis elements under two conditions: in a traditional learning paradigm with a consistent force field, and in a random sequence of force fields where learning is not possible. Remarkably, we find that the basis remains invariant. 1 Introd uction

****

A Parallel Mixture of SVMs for Very Large Scale Problems

Ronan Collobert, Samy Bengio, Yoshua Bengio

Support Vector Machines (SVMs) are currently the state-of-the-art models for many classification problems but they suffer from the complexity of their train(cid:173) ing algorithm which is at least quadratic with respect to the number of examples. Hence, it is hopeless to try to solve real-life problems havin g more than a few hundreds of thousands examples with SVMs. The present paper proposes a new mixture of SVMs that can be easily implemented in parallel and where each SVM is trained on a small subset of the whole dataset. Experiments on a large benchmark dataset (Forest) as well as a difficult speech database, yielded significant time improvement (time com plexity appears empirically to locally grow linearly with the number of exa mples) . In addition, and that is a surprise, a significant improvement in g eneralization was observed on Forest.

****

Switch Packet Arbitration via Queue-Learning

Timothy Brown

In packet switches, packets queue at switch inputs and contend for out- puts. Th e contention arbitration policy directly affects switch perfor- mance. The best policy depends on the current state of the switch and current traf■c patterns. T his problem is hard because the state space, possible transitions, and set of ac tions all grow exponentially with the size of the switch. We present a reinforce ment learning formulation of the problem that decomposes the value function into many small inde- pendent value functions and enables an ef■cient action selecti on.

****

Multi Dimensional ICA to Separate Correlated Sources

Roland Vollgraf, Klaus Obermayer

We present a new method for the blind separation of sources, which do not fulfi ll the independence assumption. In contrast to standard methods we consider groups of neighboring samples ("patches") within the observed mixtures. F irst we extract independent features from the observed patches. It turns out that the average dependencies between these features in different sourc es is in general lower than the dependencies be(cid:173) tween the amp litudes of different sources. We show that it might be the case that most of the dependencies is carried by only a small number of feature s. Is this case - provided these features can be identified by some heuristic - we project all patches into the subspace which is orthogona l to the subspace spanned by the "correlated" features. Standard ICA is th en performed on the elements of the transformed patches (for which the inde pendence assumption holds) and ro(cid:173) bustly yields a good estimate of the mixing matrix.

****

Effective Size of Receptive Fields of Inferior Temporal Visual Cortex Neurons in Natural Scenes

Thomas Trappenberg, Edmund Rolls, Simon Stringer

Inferior temporal cortex (IT) neurons have large receptive ■elds when a single e ffective object stimulus is shown against a blank background, but have much smal ler receptive ■elds when the object is placed in a natural scene. Thus, translat ion invariant object recognition is reduced in natural scenes, and this may help object selection. We describe a model which accounts for this by competition wi

thin an attractor in which the neurons are tuned to different objects in the sce ne, and the fovea has a higher cortical magni■cation factor than the peripheral visual ■eld. Further- more, we show that top-down object bias can increase the r eceptive ■eld size, facilitating object search in complex visual scenes, and pro viding a model of object-based attention. The model leads to the prediction that introduction of a second object into a scene with blank background will reduce the receptive ■eld size to values that depend on the closeness of the second obj ect to the target stimulus. We suggest that mechanisms of this type enable the o utput of IT to be primarily about one object, so that the areas that receive fro m IT can select the object as a potential target for action.
****

A kernel method for multi-labelled classification
André Elisseeff, Jason Weston
This article presents a Support Vector Machine (SVM) like learning sys- tem to h andle multi-label problems. Such problems are usually decom- posed into many two -class problems but the expressive power of such a system can be weak [5, 7]. We explore a new direct approach. It is based on a large margin ranking system tha t shares a lot of common proper- ties with SVMs. We tested it on a Yeast gene fu nctional classi■cation problem with positive results.
****

Learning Lateral Interactions for Feature Binding and Sensory Segmentation
Heiko Wersing
We present a new approach to the supervised learning of lateral inter- actions f or the competitive layer model (CLM) dynamic feature binding architecture. The m ethod is based on consistency conditions, which were recently shown to character ize the attractor states of this linear threshold recurrent network. For a given set of training examples the learning prob- lem is formulated as a convex quadr atic optimization problem in the lat- eral interaction weights. An ef■cient dime nsion reduction of the learning problem can be achieved by using a linear superp osition of basis inter- actions. We show the successful application of the metho d to a medical image segmentation problem of ■uorescence microscope cell images.
****

Very loopy belief propagation for unwrapping phase images
Brendan J. Frey, Ralf Koetter, Nemanja Petrovic
Since the discovery that the best error-correcting decoding algo(cid:173) rithm can be viewed as belief propagation in a cycle-bound graph, researche rs have been trying to determine under what circum(cid:173) stances "loo py belief propagation" is effective for probabilistic infer(cid:173) ence. Desp ite several theoretical advances in our understanding of loopy belief propagati on, to our knowledge, the only problem that has been solved using loopy belief propagation is error-correcting decoding on Gaussian channels. We pro pose a new representation for the two-dimensional phase unwrapping problem, and we show that loopy belief propagation produces results that are superio r to existing techniques. This is an important result, since many imag(cid:173) ing techniques, including magnetic resonance imaging and interfer(cid:173) o metric synthetic aperture radar, produce phase-wrapped images. Interestingly, the graph that we use has a very large number of very short cycles, supporting evidence that a large minimum cycle length is not needed for exc ellent results using belief propagation.
****

Modularity in the motor system: decomposition of muscle patterns as combinations of time-varying synergies
A. D'avella, M. Tresch
The question of whether the nervous system produces movement through the combina tion of a few discrete elements has long been central to the study of motor cont rol. Muscle synergies, i.e. coordinated patterns of muscle activity, have been p roposed as possible building blocks. Here we propose a model based on combinatio ns of muscle synergies with a spe- ci■c amplitude and temporal structure. Time-v arying synergies provide a realistic basis for the decomposition of the complex patterns observed in natural behaviors. To extract time-varying synergies from s

imultane- ous recording of EMG activity we developed an algorithm which extends existing non-negative matrix factorization techniques.
****

The Method of Quantum Clustering
David Horn, Assaf Gottlieb
We propose a novel clustering method that is an extension of ideas inher- ent to scale-space clustering and support-vector clustering. Like the lat- ter, it associates every data point with a vector in Hilbert space, and like the former it puts emphasis on their total sum, that is equal to the scale- space probability function. The novelty of our approach is the study of an operator in Hilbert space, represented by the Schr¨odinger equation of which the probability function is a solution. This Schr¨odinger equation contains a potential function that can be derived analytically from the probability function. We associate minima of the potential with cluster centers. The method has one variable parameter, the scale of its Gaussian kernel. We demonstrate its applicability on known data sets. By limiting the evaluation of the Schr¨odinger potential to the locations of data points, we can apply this method to problems in high dimensions.
****

Generating velocity tuning by asymmetric recurrent connections
Xiaohui Xie, Martin Giese
Asymmetric lateral connections are one possible mechanism that can ac- count for the direction selectivity of cortical neurons. We present a math- ematical analysis for a class of these models. Contrasting with earlier theoretical work that has relied on methods from linear systems theory, we study the network's nonlinear dynamic properties that arise when the threshold nonlinearity of the neurons is taken into account. We show that such networks have stimulus-locked traveling pulse solutions that are appropriate for modeling the responses of direction selective cortical neurons. In addition, our analysis shows that outside a certain regime of stimulus speeds the stability of this solutions breaks down giving rise to another class of solutions that are characterized by speci■c spatio- temporal periodicity. This predicts that if direction selectivity in the cor- tex is mainly achieved by asymmetric lateral connections lurching activ- ity waves might be observable in ensembles of direction selective cortical neurons within appropriate regimes of the stimulus speed.
****

Incremental Learning and Selective Sampling via Parametric Optimization Framework for SVM
Shai Fine, Katya Scheinberg
We propose a framework based on a parametric quadratic program(cid:173) ming (QP) technique to solve the support vector machine (SVM) training problem. This framework, can be specialized to obtain two SVM optimization methods. The first solves the fixed bias prob(cid:173) lem, while the second starts with an optimal solution for a fixed bias problem and adjusts the bias until the optimal value is found. The later method can be applied in conjunction with any other ex(cid:173) isting technique which obtains a fixed bias solution. Moreover, the second method can also be used independently to solve the com(cid:173) plete SVM training problem. A combination of these two methods is more flexible than each individual method and, among other things, produces an incremental algorithm which exactly solve the 1-Norm Soft Margin SVM optimization problem. Applying Selec(cid:173) tive Sampling techniques may further boost convergence.
****

Bayesian Predictive Profiles With Applications to Retail Transaction Data
Igor Cadez, Padhraic Smyth
Massive transaction data sets are recorded in a routine manner in telecommunications, retail commerce, and Web site management. In this paper we address the problem of inferring predictive in- dividual pro■les from such historical transaction data. We de- scribe a generative mixture model for count data and use an an approximate Bayesian estimation framework that e■ectively com- bines an individual's speci■c history with more general population patterns. We use a large real-w

orld retail transaction data set to illustrate how these pro■les consistently ou tperform non-mixture and non-Bayesian techniques in predicting customer behavior in out-of-sample data.

****

A Rational Analysis of Cognitive Control in a Speeded Discrimination Task
Michael C. Mozer, Michael Colagrosso, David Huber

We are interested in the mechanisms by which individuals monitor and adjust thei r performance of simple cognitive tasks. We model a speeded discrimination task in which individuals are asked to classify a sequence of stimuli (Jones & Braver , 2001). Response con■ict arises when one stimulus class is infrequent relative to another, resulting in more errors and slower reaction times for the infrequen t class. How do control pro- cesses modulate behavior based on the relative clas s frequencies? We explain performance from a rational perspective that casts the goal of individuals as minimizing a cost that depends both on error rate and re - action time. With two additional assumptions of rationality—that class prior p robabilities are accurately estimated and that inference is optimal subject to l imitations on rate of information transmission—we obtain a good ■t to overall RT and error data, as well as trial-by-trial variations in performance.

****

Learning Spike-Based Correlations and Conditional Probabilities in Silicon
Aaron Shon, David Hsu, Chris Diorio

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.

****

A Natural Policy Gradient
Sham M. Kakade

We provide a natural gradient method that represents the steepest descent dir ection based on the underlying structure of the param(cid:173) eter space. Alth ough gradient methods cannot make large changes in the values of the parame ters, we show that the natural gradi(cid:173) ent is moving toward choosin g a greedy optimal action rather than just a better action. These greedy opti mal actions are those that would be chosen under one improvement step of policy iteration with approximate, compatible value functions, as d efined by Sut(cid:173) ton et al. [9]. We then show drastic performance im provements in simple MDPs and in the more challenging MDP of Tetris.

****

Spectral Kernel Methods for Clustering
Nello Cristianini, John Shawe-Taylor, Jaz Kandola

In this paper we introduce new algorithms for unsupervised learn(cid:173) ing ba sed on the use of a kernel matrix. All the information re(cid:173) quired by suc h algorithms is contained in the eigenvectors of the matrix or of closely relat ed matrices. We use two different but re(cid:173) lated cost functions, the Alig nment and the 'cut cost'. The first one is discussed in a companion paper [3], the second one is based on graph theoretic concepts. Both functions measure the level of clustering of a labeled dataset, or the correlation between data clus (cid:173) ters and labels. We state the problem of unsupervised learning as ass igning labels so as to optimize these cost functions. We show how the optimal s olution can be approximated by slightly relaxing the corresponding optimization problem, and how this corresponds to using eigenvector information. The result ing simple algorithms are tested on real world data with positive results.

****

Batch Value Function Approximation via Support Vectors
Thomas Dietterich, Xin Wang

We present three ways of combining linear programming with the kernel trick to f ind value function approximations for reinforcement learning. One formulation is based on SVM regression; the second is based on the Bellman equation; and the t hird seeks only to ensure that good moves have an advantage over bad moves. All formu(cid:173) lations attempt to minimize the number of support vectors while f

itting the data. Experiments in a difficult, synthetic maze problem show that al l three formulations give excellent performance, but the advantage formulation i s much easier to train. Unlike policy gradi(cid:173) ent methods, the kernel met hods described here can easily 'adjust the complexity of the function approximat or to fit the complexity of the value function.

****

PAC Generalization Bounds for Co-training

Sanjoy Dasgupta, Michael Littman, David McAllester

The rule-based bootstrapping introduced by Yarowsky, and its co- training varian t by Blum and Mitchell, have met with considerable em- pirical success. Earlier work on the theory of co-training has been only loosely related to empirically u seful co-training algorithms. Here we give a new PAC-style bound on generalizati on error which justi■es both the use of con■dences — partial rules and partial l abeling of the unlabeled data — and the use of an agreement-based objective func tion as sug- gested by Collins and Singer. Our bounds apply to the multiclass ca se, i.e., where instances are to be assigned one of

****

Face Recognition Using Kernel Methods

Ming-Hsuan Yang

Principal Component Analysis and Fisher Linear Discriminant methods have demonst rated their success in face detection, recog(cid:173) nition, and tracking. The representation in these subspace methods is based on second order statistics of the image set, and does not address higher order statistical dependencies such a s the relation(cid:173) ships among three or more pixels. Recently Higher Order Statistics and Independent Component Analysis (ICA) have been used as in(cid:173) formative low dimensional representations for visual recognition. In this pape r, we investigate the use of Kernel Principal Compo(cid:173) nent Analysis and K ernel Fisher Linear Discriminant for learning low dimensional representations fo r face recognition, which we call Kernel Eigenface and Kernel Fisherface methods . While Eigenface and Fisherface methods aim to find projection directions based on the second order correlation of samples, Kernel Eigenface and Ker(cid:173) n el Fisherface methods provide generalizations which take higher order correlatio ns into account. We compare the performance of kernel methods with Eigenface, Fi sherface and ICA-based meth(cid:173) ods for face recognition with variation in pose, scale, lighting and expression. Experimental results show that kernel meth ods pro(cid:173) vide better representations and achieve lower error rates for f ace recognition.

****

Active Portfolio-Management based on Error Correction Neural Networks

Hans-Georg Zimmermann, Ralph Neuneier, Ralph Grothmann

This paper deals with a neural network architecture which establishes a portfoli o management system similar to the Black / Litterman approach. This allocation s cheme distributes funds across various securities or ■- nancial markets while si multaneously complying with speci■c allocation constraints which meet the requir ements of an investor. The portfolio optimization algorithm is modeled by a feed forward neural network. The underlying expected return forecasts are based on er ror correction neural networks (ECNN), which utilize the last model error as an auxiliary input to evaluate their own misspeci■cation. The portfolio optimizatio n is implemented such that (i.) the allocations comply with investor’s constrain ts and that (ii.) the risk of the portfo- lio can be controlled. We demonstrate the pro■tability of our approach by constructing internationally diversi■ed port folios across different ■nancial markets of the G7 contries. It turns out, that our approach is superior to a preset benchmark portfolio.

****

Generalization Performance of Some Learning Problems in Hilbert Functional Space s

T. Zhang

We investigate the generalization performance of some learning prob- lems in Hil bert functional Spaces. We introduce a notion of convergence of the estimated fu nctional predictor to the best underlying predictor, and obtain an estimate on t

he rate of the convergence. This estimate allows us to derive generalization bou nds on some learning formulations.
****

Analog Soft-Pattern-Matching Classifier using Floating-Gate MOS Technology
Toshihiko Yamasaki, Tadashi Shibata
A flexible pattern-matching analog classifier is presented in con- junction with a robust image representation algorithm called Prin- cipal Axes Projection (PAP). In the circuit, the functional form of matching is configurable in terms of the peak position, the peak height and the sharpness of the similarity evaluati on. The test chip was fabri- cated in a 0.6-m m CMOS technology and successfully applied to hand-written pattern recognition and medical radiograph analysis us ing PAP as a feature extraction pre-processing step for robust image coding. T he separation and classification of overlapping patterns is also ex- perimentall y demonstrated.
****

Generalizable Relational Binding from Coarse-coded Distributed Representations
Randall O'Reilly, R. Busby
We present a model of binding of relationship information in a spatial domain (e .g., square above triangle) that uses low-order coarse-coded
****

On the Concentration of Spectral Properties
John Shawe-Taylor, Nello Cristianini, Jaz Kandola
We consider the problem of measuring the eigenvalues of a ran(cid:173) domly drawn sample of points. We show that these values can be reliably e stimated as can the sum of the tail of eigenvalues. Fur(cid:173)thermore , the residuals when data is projected into a subspace is shown to be r eliably estimated on a random sample. Experiments are presented that confirm the theoretical results.
****

Small-World Phenomena and the Dynamics of Information
Jon Kleinberg
The problem of searching for information in networks like the World Wide Web can be approached in a variety of ways, ranging from centralized indexing schemes to decentralized mechanisms that navigate the underlying network without knowledge of its global structure. The decentralized approach appears in a variety of sett ings:
in the behavior of users browsing the Web by following hyperlinks; in the design of
****

Stochastic Mixed-Signal VLSI Architecture for High-Dimensional Kernel Machines
Roman Genov, Gert Cauwenberghs
A mixed-signal paradigm is presented for high-resolution parallel inner- product computation in very high dimensions, suitable for ef■cient im- plementation of kernels in image processing. At the core of the externally digital architecture is a high-density, low-power analog array performing binary-binary partial matri x-vector multiplication. Full digital resolution is maintained even with low-res olution analog-to-digital conversion, ow- ing to random statistics in the analog summation of binary products. A random modulation scheme produces near-Bernoull i statistics even for highly correlated inputs. The approach is validated with r eal image data, and with experimental results from a CID/DRAM analog array proto type in 0.5
****

The Emergence of Multiple Movement Units in the Presence of Noise and Feedback D elay
Michael Kositsky, Andrew Barto
Tangential hand velocity pro■les of rapid human arm movements of- ten appear as sequences of several bell-shaped acceleration-deceleration phases called submove ments or movement units. This suggests how the nervous system might ef■ciently c ontrol a motor plant in the presence of noise and feedback delay. Another critic al observation is that stochastic- ity in a motor control problem makes the opti

mal control policy essen- tially different from the optimal control policy for t he deterministic case. We use a simpli■ed dynamic model of an arm and address ra pid aimed arm movements. We use reinforcement learning as a tool to approximate the optimal policy in the presence of noise and feedback delay. Using a simpli■e d model we show that multiple submovements emerge as an optimal policy in the pr esence of noise and feedback delay. The optimal policy in this situation is to d rive the arm’s end point close to the target by one fast submovement and then ap ply a few slow submovements to accu- rately drive the arm’s end point into the t arget region. In our simulations, the controller sometimes generates corrective submovements before the initial fast submovement is completed, much like the pre dictive correc- tions observed in a number of psychophysical experiments.
****

Pranking with Ranking
Koby Crammer, Yoram Singer
We discuss the problem of ranking instances. In our framework each ins tance is associated with a rank or a rating, which is an integer fro m 1 to k. Our goal is to find a rank-prediction rule that assigns each i nstance a rank which is as close as possible to the instance's true r ank. We describe a simple and efficient online al(cid:173) gorithm, analy ze its performance in the mistake bound model, and prove its correctness. We describe two sets of experiments, with synthetic data and with the Ea chMovie dataset for collaborative filtering. In the experiments we perfor med, our algorithm outper(cid:173) forms online algorithms for regression and classification applied to ranking.
****

Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning
Evan Greensmith, Peter Bartlett, Jonathan Baxter
We consider the use of two additive control variate methods to reduce the varian ce of performance gradient estimates in reinforcement learn- ing problems. The ■ rst approach we consider is the baseline method, in which a function of the curr ent state is added to the discounted value estimate. We relate the performance o f these methods, which use sam- ple paths, to the variance of estimates based on iid data. We derive the baseline function that minimizes this variance, and we show that the vari- ance for any baseline is the sum of the optimal variance and a weighted squared distance to the optimal baseline. We show that the widely us ed average discounted value baseline (where the reward is replaced by the differ ence between the reward and its expectation) is suboptimal. The second approach we consider is the actor-critic method, which uses an approximate value function . We give bounds on the expected squared error of its estimates. We show that mi nimizing distance to the true value function is suboptimal in general; we provid e an example for which the true value function gives an estimate with positive v ariance, but the op- timal value function gives an unbiased estimate with zero v ariance. Our bounds suggest algorithms to estimate the gradient of the performan ce of parameterized baseline or value functions. We present preliminary exper- i ments that illustrate the performance improvements on a simple control problem.
****

Asymptotic Universality for Learning Curves of Support Vector Machines
Manfred Opper, Robert Urbanczik
Using methods of Statistical Physics, we investigate the rOle of model complexity in learning with support vector machines (SVMs). We show the advantages of using SVMs with kernels of infinite complexity on no isy target rules, which, in contrast to common theoretical beliefs, ar e found to achieve optimal general(cid:173) ization error although the train ing error does not converge to the generalization error. Moreover, we fin d a universal asymptotics of the learning curves which only depend on t he target rule but not on the SVM kernel.
****

Active Information Retrieval
Tommi Jaakkola, Hava Siegelmann
In classical large information retrieval systems, the system responds to a use

r initiated query with a list of results ranked by relevance. The users may further refine their query as needed. This process may result in a l engthy correspondence without conclusion. We propose an alternative acti ve learning approach, where the sys(cid:173) tem responds to the initial us er's query by successively probing the user for distinctions at multiple level s of abstraction. The system's initiated queries are optimized for speedy recovery and the user is permitted to respond with multiple selections o r may reject the query. The information is in each case unambiguously incorpor ated by the system and the subsequent queries are adjusted to minimize the ne ed for further exchange. The system's initiated queries are subject to resource constraints pertaining to the amount of infor(cid:173) mation that ca n be presented to the user per iteration.

****

Tempo tracking and rhythm quantization by sequential Monte Carlo

Ali Taylan Cemgil, Bert Kappen

We present a probabilistic generative model for timing deviations in expressive music. performance. The structure of the proposed model is equivalent to a switc hing state space model. We formu(cid:173) late two well known music recognition problems, namely tempo tracking and automatic transcription (rhythm quantization) as fil(cid:173) tering and maximum a posteriori (MAP) state estimation tasks. The inferences are carried out using sequential Monte Carlo in(cid:173) tegratio n (particle filtering) techniques. For this purpose, we have derived a novel Vit erbi algorithm for Rao-Blackwellized particle fil(cid:173) ters, where a subset of the hidden variables is integrated out. The resulting model is suitable for r ealtime tempo tracking and tran(cid:173) scription and hence useful in a number of music applications such as adaptive automatic accompaniment and score typeset ting.

****

Learning a Gaussian Process Prior for Automatically Generating Music Playlists

John Platt, Christopher J. C. Burges, Steven Swenson, Christopher Weare, Alice Z heng

This paper presents AutoDJ: a system for automatically generating mu- sic playli sts based on one or more seed songs selected by a user. AutoDJ uses Gaussian Pro cess Regression to learn a user preference function over songs. This function ta kes music metadata as inputs. This paper further introduces Kernel Meta-Training , which is a method of learning a Gaussian Process kernel from a distribution of functions that generates the learned function. For playlist generation, AutoDJ learns a kernel from a large set of albums. This learned kernel is shown to be m ore effective at predicting users' playlists than a reasonable hand-designed ker nel.

****

Exact differential equation population dynamics for integrate-and-fire neurons

Julian Eggert, Berthold Bäuml

In our previous work, integral equation formulations for

****

Improvisation and Learning

Judy A. Franklin

This article presents a 2-phase computational learning model and appli- cation. As a demonstration, a system has been built, called CHIME for Computer Human Int eracting Musical Entity. In phase 1 of training, re- current back-propagationtra ins the machine to reproduce 3 jazz melodies. The recurrent network is expanded and is further trained in phase 2 with a reinforcement learning algorithm and a critique produced by a set of basic rules for jazz improvisation. After each pha se CHIME can interactively improvise with a human in real time.

****

Geometrical Singularities in the Neuromanifold of Multilayer Perceptrons

Shun-ichi Amari, Hyeyoung Park, Tomoko Ozeki

Singularities are ubiquitous in the parameter space of hierarchical models suc h as multilayer perceptrons. At singularities, the Fisher information matrix degenerates, and the Cramer-Rao paradigm does no more hold, implying that

the classical model selection the(cid:173) ory such as AIC and MDL cannot be applied. It is important to study the relation between the generalizati on error and the training error at singularities. The present paper demon strates a method of analyzing these errors both for the maximum likelihood e stima(cid:173) tor and the Bayesian predictive distribution in terms of Gau ssian random fields, by using simple models.

****

A Sequence Kernel and its Application to Speaker Recognition

William Campbell

A novel approach for comparing sequences of observations using an explicit-expan sion kernel is demonstrated. The kernel is derived using the assumption of the i ndependence of the sequence of observations and a mean-squared error training cr iterion. The use of an explicit expan- sion kernel reduces classi■er model size and computation dramatically, resulting in model sizes and computation one-hundr ed times smaller in our application. The explicit expansion also preserves the c omputational advantages of an earlier architecture based on mean-squared error t rain- ing. Training using standard support vector machine methodology gives accu racy that signi■cantly exceeds the performance of state-of-the-art mean-squared error training for a speaker recognition task.

****

Convergence of Optimistic and Incremental Q-Learning

Eyal Even-dar, Yishay Mansour

Vie sho,v the convergence of tV/O deterministic variants of Q(cid:173) learning. The first is the widely used optimistic Q-learning, which initializes the Q-val ues to large initial values and then follows a greedy policy with respect to the Q-values. We show that setting the initial value sufficiently large guarantees the converges to an E(cid:173) optimal policy. The second is a new and novel alg orithm incremen(cid:173) tal Q-learning, which gradually promotes the values of actions that are not taken. We show that incremental Q-learning converges, in th e limit, to the optimal policy. Our incremental Q-learning algo(cid:173) rithm c an be viewed as derandomization of the E-greedy Q-learning.

****

Rao-Blackwellised Particle Filtering via Data Augmentation

Christophe Andrieu, Nando Freitas, Arnaud Doucet

EE Engineering

****

Boosting and Maximum Likelihood for Exponential Models

Guy Lebanon, John Lafferty

We derive an equivalence between AdaBoost and the dual of a convex optimization problem, showing that the only difference between mini- mizing the exponential l oss used by AdaBoost and maximum likelihood for exponential models is that the l atter requires the model to be normal- ized to form a conditional probability di stribution over labels. In addi- tion to establishing a simple and easily unders tood connection between the two methods, this framework enables us to derive new regularization procedures for boosting that directly correspond to penalized ma ximum likelihood. Experiments on UCI datasets support our theoretical analy- sis and give additional insight into the relationship between boosting and logistic regression.

****

Iterative Double Clustering for Unsupervised and Semi-Supervised Learning

Ran El-Yaniv, Oren Souroujon

We present a powerful meta-clustering technique called Iterative Dou- ble Cluste ring (IDC). The IDC method is a natural extension of the recent Double Clusterin g (DC) method of Slonim and Tishby that ex- hibited impressive performance on te xt categorization tasks [12]. Us- ing synthetically generated data we empiricall y ■nd that whenever the DC procedure is successful in recovering some of the str ucture hidden in the data, the extended IDC procedure can incrementally compute a signi■cantly more accurate classi■cation. IDC is especially advan- tageous whe n the data exhibits high attribute noise. Our simulation results also show the e ■ectiveness of IDC in text categorization prob- lems. Surprisingly, this unsuper

vised procedure can be competitive with a (supervised) SVM trained with a small training set. Finally, we propose a simple and natural extension of IDC for semi-supervised and transductive learning where we are given both labeled and unlabeled examples.

****

Multiagent Planning with Factored MDPs

Carlos Guestrin, Daphne Koller, Ronald Parr

We present a principled and ef■cient planning algorithm for cooperative multiagent dynamic systems. A striking feature of our method is that the coordination and communication between the agents is not imposed, but derived directly from the system dynamics and function approximation architecture. We view the en- tire multiagent system as a single, large Markov decision process (MDP), which we assume can be represented in a factored way using a dynamic Bayesian net- work (DBN). The action space of the resulting MDP is the joint action space of the entire set of agents. Our approach is based on the use of factored linear value functions as an approximation to the joint value function. This factorization of the value function allows the agents to coordinate their actions at runtime using a natural message passing scheme. We provide a simple and ef■cient method for computing such an approximate value function by solving a single linear pro- gram, whose size is determined by the interaction between the value function structure and the DBN. We thereby avoid the exponential blowup in the state and action space. We show that our approach compares favorably with approaches based on reward sharing. We also show that our algorithm is an ef■cient alterna- tive to more complicated algorithms even in the single agent case.

****

ACh, Uncertainty, and Cortical Inference

Peter Dayan, Angela J. Yu

Acetylcholine (ACh) has been implicated in a wide variety of tasks involving attentional processes and plasticity. Following extensive animal studies, it has previously been suggested that ACh reports on uncertainty and controls hippocampal, cortical and cortico-amygdalar plasticity. We extend this view and consider its effects on cortical representational inference, arguing that ACh controls the balance between bottom-up inference, in(cid:3)uenced by input stimuli, and top-down inference, in(cid:3)uenced by contextual information. We illustrate our proposal using a hierarchical hid- den Markov model.

****

On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes

Andrew Ng, Michael Jordan

We compare discriminative and generative learning as typified by logistic regression and naive Bayes. We show, contrary to a widely(cid:173) held belief that discriminative classifiers are almost always to be preferred, that there can often be two distinct regimes of per(cid:173) formance as the training set size is increased, one in which each algorithm does better. This stems from the observation- which is borne out in repeated experiments- that while discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster.

****

Risk Sensitive Particle Filters

Sebastian Thrun, John Langford, Vandi Verma

We propose a new particle ■lter that incorporates a model of costs when generating particles. The approach is motivated by the observation that the costs of accidentally not tracking hypotheses might be signi■cant in some areas of state space, and next to irrelevant in others. By incorporat- ing a cost model into particle ■ltering, states that are more critical to the system performance are more likely to be tracked. Automatic calculation of the cost model is implemented using an MDP value function calcula- tion that estimates the value of tracking a particular state. Experiments in two mobile robot domains illustrate the appropriateness of the approach.

****
Characterizing Neural Gain Control using Spike-triggered Covariance
Odelia Schwartz, E.J. Chichilnisky, Eero Simoncelli
Spike-triggered averaging techniques are effective for linear characterization o f neural responses. But neurons exhibit important nonlinear behaviors, such as g ain control, that are not captured by such analyses. We describe a spike-trigger ed covariance method for retrieving suppressive components of the gain control s ignal in a neuron. We demonstrate the method in simulation and on retinal gangli on cell data. Analysis of physiological data reveals significant suppressive axe s and explains neural nonlinearities. This method should be applicable to other sensory areas and modalities.
****
Speech Recognition with Missing Data using Recurrent Neural Nets
S. Parveen, P. Green
In the missing data' approach to improving the robustness of automatic speech re cognition to added noise, an initial process identifies spectral- temporal regio ns which are dominated by the speech source. The remaining regions are considere d to bemissing'. In this paper we develop a connectionist approach to the proble m of adapting speech recognition to the missing data case, using Recurrent Neura l Networks. In contrast to methods based on Hidden Markov Models, RNNs allow us to make use of long-term time constraints and to make the problems of classifica tion with incomplete data and imputing missing values interact. We report encour aging results on an isolated digit recognition task. 1. Introduction
****
On Spectral Clustering: Analysis and an algorithm
Andrew Ng, Michael Jordan, Yair Weiss
Despite many empirical successes of spectral clustering methods(cid:173) algor ithms that cluster points using eigenvectors of matrices de(cid:173) riv ed from the data- there are several unresolved issues. First, there ar e a wide variety of algorithms that use the eigenvectors in slightly different ways. Second, many of these algorithms have no proof that the y will actually compute a reasonable clustering. In this paper, we pr esent a simple spectral clustering algorithm that can be implemented using a few lines of Matlab. Using tools from matrix perturbation theory, w e analyze the algorithm, and give conditions under which it can be e xpected to do well. We also show surprisingly good experimental result s on a number of challenging clustering problems.
****
Grouping with Bias
Stella X. Yu, Jianbo Shi
With the optimization of pattern discrimination as a goal, graph partitioning ap proaches often lack the capability to integrate prior knowledge to guide groupin g. In this paper, we consider priors from unitary generative models, partially l abeled data and spatial attention. These priors are modelled as constraints in t he solution space. By imposing uniformity condition on the constraints, we restr ict the feasible space to one of smooth solutions. A subspace projection method is developed to solve this constrained eigenprob(cid:173) lema We demonstrate th at simple priors can greatly improve image segmentation results.
****
Learning Hierarchical Structures with Linear Relational Embedding
Alberto Paccanaro, Geoffrey E. Hinton
We present Linear Relational Embedding (LRE), a new method of learn- ing a distr ibuted representation of concepts from data consisting of in- stances of relatio ns between given concepts. Its ■nal goal is to be able to generalize, i.e. infer new instances of these relations among the con- cepts. On a task involving fami ly relationships we show that LRE can generalize better than any previously publ ished method. We then show how LRE can be used effectively to ■nd compact distri buted representa- tions for variable-sized recursive data structures, such as tr ees and lists.
****

Constructing Distributed Representations Using Additive Clustering
Wheeler Ruml
If the promise of computational modeling is to be fully realized in higher- level cognitive domains such as language processing, principled methods must be developed to construct the semantic representations used in such models. In this paper, we propose the use of an established formalism from mathematical psychology, additive clustering, as a means of auto- matically constructing binary representations for objects using only pair- wise similarity data. However, existing methods for the unsupervised learning of additive clustering models do not scale well to large prob- lems. We present a new algorithm for additive clustering, based on a novel heuristic technique for combinatorial optimization. The algorithm is simpler than previous formulations and makes fewer independence as- sumptions. Extensive empirical tests on both human and synthetic data suggest that it is more effective than previous methods and that it also scales better to larger problems. By making additive clustering practical, we take a significant step toward scaling connectionist models beyond hand-coded examples. 1 Introduction
****

Products of Gaussians
Christopher Williams, Felix Agakov, Stephen Felderhof
Recently Hinton (1999) has introduced the Products of Experts (PoE) model in which several individual probabilistic models for data are combined to provide an overall model of the data. Be(cid:173)low we consider PoE models in which each expert is a Gaussian. Although the product of Gaussians is also a Gaussian, if each Gaus(cid:173)sian has a simple structure the product can have a richer structure. We examine (1) Products of Gaussian pancakes which give rise to proba bilistic Minor Components Analysis, (2) products of I-factor PPCA models and (3) a products of experts construction for an AR(l) process.
****

The Fidelity of Local Ordinal Encoding
Javid Sadr, Sayan Mukherjee, Keith Thoresz, Pawan Sinha
A key question in neuroscience is how to encode sensory stimuli such as images and sounds. Motivated by studies of response prop- erties of neurons in the early cortical areas, we propose an encoding scheme that dispenses with absolute measures of signal intensity or contrast and uses, instead, only local ordinal measures. In this scheme, the structure of a signal is represented by a set of equalities and inequalities across adjacent regions. In this paper, we focus on characterizing the (cid:12)delity of this representation strategy. We develop a regularization approach for image reconstruction from ordinal measures and thereby demonstrate that the ordinal repre- sentation scheme can faithfully encode signal structure. We also present a neurally plausible implementation of this computation that uses only local update rules. The results highlight the robust- ness and generalization ability of local ordinal encodings for the task of pattern classi(cid:12)cation.
****

Global Coordination of Local Linear Models
Sam Roweis, Lawrence Saul, Geoffrey E. Hinton
High dimensional data that lies on or near a low dimensional manifold can be de- scribed by a collection of local linear models. Such a description, however, does not provide a global parameterization of the manifold—arguably an important goal of unsupervised learning. In this paper, we show how to learn a collection of local linear models that solves this more dif■cult problem. Our local linear models are represented by a mixture of factor analyzers, and the “global coordi- nation” of these models is achieved by adding a regularizing term to the standard maximum likelihood objective function. The regularizer breaks a degeneracy in the mixture model’s parameter space, favoring models whose internal coor- dinate systems are aligned in a consistent way. As a result, the internal coor- dinates change smoothly and continuously as one traverses a connected path on the manifold—even when the path crosses the domains of many different local models. The regularizer takes the form of a Kullback-Leibler divergence and illustrates an unexpected application of variational methods: not to perform ap- proximate infer

ence in intractable probabilistic models, but to learn more useful internal repr
esentations in tractable ones.
****

Correlation Codes in Neuronal Populations
Maoz Shamir, Haim Sompolinsky
Population codes often rely on the tuning of the mean responses to the stimulus
parameters. However, this information can be greatly sup- pressed by long range
correlations. Here we study the ef■ciency of cod- ing information in the second
order statistics of the population responses. We show that the Fisher Informatio
n of this system grows linearly with the size of the system. We propose a biline
ar readout model for extract- ing information from correlation codes, and evalua
te its performance in discrimination and estimation tasks. It is shown that the
main source of information in this system is the stimulus dependence of the vari
ances of the single neuron responses.
****

Grammatical Bigrams
Mark Paskin
Unsupervised learning algorithms have been derived for several sta(cid:173) tist
ical models of English grammar, but their computational com(cid:173) plexity
makes applying them to large data sets intractable. This paper presents
a probabilistic model of English grammar that is much simpler than con
ventional models, but which admits an effi(cid:173) cient EM training algorit
hm. The model is based upon grammat(cid:173) ical bigrams, i.e. , syntact
ic relationships between pairs of words. We present the results of exper
iments that quantify the represen(cid:173) tational adequacy of the grammatic
al bigram model, its ability to generalize from labelled data, and its ab
ility to induce syntactic structure from large amounts of raw text.
****

The Steering Approach for Multi-Criteria Reinforcement Learning
Shie Mannor, Nahum Shimkin
We consider the problem of learning to attain multiple goals in a dynamic envi-
ronment, which is initially unknown. In addition, the environment may contain ar
bitrarily varying elements related to actions of other agents or to non-stationa
ry moves of Nature. This problem is modelled as a stochastic (Markov) game betwe
en the learning agent and an arbitrary player, with a vector-valued reward funct
ion. The objective of the learning agent is to have its long-term average reward
vector belong to a given target set. We devise an algorithm for achieving this
task, which is based on the theory of approachability for stochastic games. This
algorithm com- bines, in an appropriate way, a ■nite set of standard, scalar-re
ward learning algo- rithms. Su–cient conditions are given for the convergence of
the learning algorithm to a general target set. The specialization of these res
ults to the single-controller Markov decision problem are discussed as well.
****

Stabilizing Value Function Approximation with the BFBP Algorithm
Xin Wang, Thomas Dietterich
We address the problem of non-convergence of online reinforcement learning algor
ithms (e.g., Q learning and SARSA(A)) by adopt(cid:173) ing an incremental-batch
approach that separates the exploration process from the function fitting proce
ss. Our BFBP (Batch Fit to Best Paths) algorithm alternates between an explorati
on phase (during which trajectories are generated to try to find fragments of th
e optimal policy) and a function fitting phase (during which a function approxim
ator is fit to the best known paths from start states to terminal states). An ad
vantage of this approach is that batch value-function fitting is a global proces
s, which allows it to address the tradeoffs in function approximation that canno
t be handled by local, online algorithms. This approach was pioneered by Boyan a
nd Moore with their GROWSUPPORT and ROUT al(cid:173) gorithms. We show how to im
prove upon their work by applying a better exploration process and by enriching
the function fitting procedure to incorporate Bellman error and advantage error
mea(cid:173) sures into the objective function. The results show improved per(ci
d:173) formance on several benchmark problems.

****
Tree-based reparameterization for approximate inference on loopy graphs
Martin J. Wainwright, Tommi Jaakkola, Alan Willsky
We develop a tree-based reparameterization framework that pro(cid:173) vides a new conceptual view of a large class of iterative algorithms for comput ing approximate marginals in graphs with cycles. It includes belief pro pagation (BP), which can be reformulated as a very local form of reparam eterization. More generally, we consider algorithms that perform exact co mputations over spanning trees of the full graph. On the practical si de, we find that such tree reparameterization (TRP) algorithms have conve rgence properties superior to BP. The reparameterization perspective also provides a number of theoretical insights into approximate inference, in(cid:173) cluding a new characterization of fixed points; and an invar iance intrinsic to TRP /BP. These two properties enable us to analyze and bound the error between the TRP /BP approximations and the actual marginals. While our results arise naturally from the TRP perspective, most of them apply in an algorithm-independent manner to any local minimum of the Bethe free energy. Our re(cid:173) sults also have natural ext ensions to more structured approxima(cid:173) tions [e.g. , 1, 2].
****
Cobot: A Social Reinforcement Learning Agent
Charles Isbell, Christian Shelton
We report on the use of reinforcement learning with Cobot, a software agent resi ding in the well-known online community LambdaMOO. Our initial work on Cobot (Is bell et al.2000) provided him with the ability to collect social statistics and report them to users. Here we describe an application of RL allowing Cobot to ta ke proactive actions in this complex social environment, and adapt behavior from multiple sources of human reward. After 5 months of training, and 3171 reward a nd punishment events from 254 different LambdaMOO users, Cobot learned nontrivia l preferences for a number of users, modi■ng his behavior based on his current s tate. Here we describe LambdaMOO and the state and action spaces of Cobot, and r eport the statistical results of the learning experiment.
****
Semi-supervised MarginBoost
Florence d'Alché-Buc, Yves Grandvalet, Christophe Ambroise
In many discrimination problems a large amount of data is available but only a few of them are labeled. This provides a strong motivation to improve or develo p methods for semi-supervised learning. In this paper, boosting is generalized to this task within the optimization framework of MarginBoost . We extend the m argin definition to unlabeled data and develop the gradient descent algorithm t hat corresponds to the resulting margin cost function. This meta-learning schem e can be applied to any base classifier able to benefit from unlabeled data. We propose here to apply it to mixture models trained with an Expectation-Maximiz ation algorithm. Promising results are presented on benchmarks with different rates of labeled data.
****
ALGONQUIN - Learning Dynamic Noise Models From Noisy Speech for Robust Speech Re cognition
Brendan J. Frey, Trausti Kristjansson, Li Deng, Alex Acero
A challenging, unsolved problem in the speech recognition com(cid:173) m unity is recognizing speech signals that are corrupted by loud, highly nonstationary noise. One approach to noisy speech recog(cid:173) nition i s to automatically remove the noise from the cepstrum se(cid:173) quence bef ore feeding it in to a clean speech recognizer. In previous work published in Eurospeech, we showed how a probability model trained on clean speech a nd a separate probability model trained on noise could be combined for the purpose of estimating the noise(cid:173) free speech from the noisy speech. We showed how an iterative 2nd order vector Taylor series approximation could be used for prob(cid:173) abilistic inference in this model. In many circumstances, it is not possible to obtain examples of noise without spe

ech. Noise statis(cid:173) tics may change significantly during an utterance , so that speech(cid:173) free frames are not sufficient for estimating the n oise model. In this paper, we show how the noise model can be learned even wh en the data contains speech. In particular, the noise model can be learned fr om the test utterance and then used to de noise the test utterance. The approxi mate inference technique is used as an approximate E step in a generalized EM algorithm that learns the parameters of the noise model from a test ut terance. For both Wall Street J our(cid:173) nal data with added noise sample s and the Aurora benchmark, we show that the new noise adaptive technique perfo rms as well as or significantly better than the non-adaptive algorithm, without the need for a separate training set of noise examples.

****

Adaptive Nearest Neighbor Classification Using Support Vector Machines

Carlotta Domeniconi, Dimitrios Gunopulos

The nearest neighbor technique is a simple and appealing method to address classification problems. It relies on t he assumption of locally const ant class conditional probabilities. This assumption becomes invalid in high dimensions with a finite number of exam(cid:173) ples due to the cur se of dimensionality. We propose a technique that computes a locally flex ible metric by means of Support Vector Machines (SVMs). The maximum margin b oundary found by the SVM is used to determine the most discriminant direction over the query's neighborhood. Such direction provides a local weighting scheme for input features. We present experimental evidence of classi fication performance improvement over the SVM algorithm alone and over a variety of adaptive learning schemes, by using both simulated and real data sets.

****

Fast, Large-Scale Transformation-Invariant Clustering

Brendan J. Frey, Nebojsa Jojic

In previous work on transformed mixtures of Gaussians'' andtransformed hidden Ma rkov models'', we showed how the EM al- gorithm in a discrete latent variable mo del can be used to jointly normalize data (e.g., center images, pitch-normalize spectrograms) and learn a mixture model of the normalized data. The only input t o the algorithm is the data, a list of possible transformations, and the number of clusters to find. The main criticism of this work was that the exhaustive com putation of the posterior probabili- ties over transformations would make scalin g up to large feature vectors and large sets of transformations intractable. Her e, we de- scribe how a tremendous speed-up is acheived through the use of a vari ational technique for decoupling transformations, and a fast Fourier transform m ethod for computing posterior probabilities. For NN images, learning C clusters under N rotations, N scales,

****

A Rotation and Translation Invariant Discrete Saliency Network

Lance Williams, John W. Zweck

We describe a neural network which enhances and completes salient closed contour s. Our work is different from all previous work in three important ways. First, like the input provided to V1 by LGN, the in- put to our computation is isotropi c. That is, the input is composed of spots not edges. Second, our network comput es a well de■ned function of the input based on a distribution of closed contour s characterized by a random process. Third, even though our computation is imple mented in a discrete network, its output is invariant to continuous rotations an d translations of the input pattern.

****

A theory of neural integration in the head-direction system

Richard Hahnloser, Xiaohui Xie, H. Seung

Integration in the head-direction system is a computation by which hor- izontal angular head velocity signals from the vestibular nuclei are in- tegrated to yie ld a neural representation of head direction. In the thala- mus, the postsubicul um and the mammillary nuclei, the head-direction representation has the form of a place code: neurons have a preferred head direction in which their ■ring is ma

ximal [Blair and Sharp, 1995, Blair et al., 1998, ?]. Integration is a dif■cult computation, given that head-velocities can vary over a large range. Previous mo dels of the head-direction system relied on the assumption that the integration is achieved in a ■ring-rate-based attractor network with a ring structure. In or der to correctly integrate head-velocity signals during high-speed head rotation s, very fast synaptic dynamics had to be assumed. Here we address the question w hether integration in the head-direction system is possible with slow synapses, for example excitatory NMDA and inhibitory GABA(B) type synapses. For neural net works with such slow synapses, rate-based dynamics are a good approximation of s pik- ing neurons [Ermentrout, 1994]. We ■nd that correct integration during high -speed head rotations imposes strong constraints on possible net- work architect ures.

****

(Not) Bounding the True Error

John Langford, Rich Caruana

We present a new approach to bounding the true error rate of a continuous valued classi■er based upon PAC-Bayes bounds. The method ■rst con- structs a distribut ion over classi■ers by determining how sensitive each parameter in the model is to noise. The true error rate of the stochastic classi■er found with the sensiti vity analysis can then be tightly bounded using a PAC-Bayes bound. In this paper we demonstrate the method on arti■cial neural networks with results of a order of magnitude im- provement vs. the best deterministic neural net bounds.

****

Novel iteration schemes for the Cluster Variation Method

Hilbert Kappen, Wim Wiegerinck

The Cluster Variation method is a class of approximation meth(cid:173) o ds containing the Bethe and Kikuchi approximations as special cases. W e derive two novel iteration schemes for the Cluster Vari(cid:173) ation M ethod. One is a fixed point iteration scheme which gives a significant impr ovement over loopy BP, mean field and TAP meth(cid:173) ods on directed grap hical models. The other is a gradient based method, that is guaranteed to converge and is shown to give useful results on random graphs with mild fru stration. We conclude that the methods are of significant practical val ue for large inference problems.

****

Infinite Mixtures of Gaussian Process Experts

Carl Rasmussen, Zoubin Ghahramani

We present an extension to the Mixture of Experts (ME) model, where the individu al experts are Gaussian Process (GP) regression models. Us- ing an input-depende nt adaptation of the Dirichlet Process, we imple- ment a gating network for an i n■nite number of Experts. Inference in this model may be done ef■ciently using a Markov Chain relying on Gibbs sampling. The model allows the effective covarian ce function to vary with the inputs, and may handle large datasets – thus potent ially over- coming two of the biggest hurdles with GP models. Simulations show t he viability of this approach.

****

Relative Density Nets: A New Way to Combine Backpropagation with HMM's

Andrew Brown, Geoffrey E. Hinton

Logistic units in the first hidden layer of a feedforward neural net(cid:173) work compute the relative probability of a data point under two Gaussia ns. This leads us to consider substituting other density models. We p resent an architecture for performing discriminative learning of Hidden Marko v Models using a network of many small HMM's. Experiments on speech data sho w it to be superior to the standard method of discriminatively training HMM's.

****

Neural Implementation of Bayesian Inference in Population Codes

Si Wu, Shun-ichi Amari

This study investigates a population decoding paradigm, in which the estimat ion of stimulus in the previous step is used as prior knowledge for c onsecutive decoding. We analyze the decoding accu(cid:173) racy of such a Baye

sian decoder (Maximum a Posteriori Estimate), and show that it can be implemented by a biologically plausible recurrent network, where the prior knowledge of stimulus is con(cid:173)veyed by the change in recurrent interactions as a result of Hebbian learning.

****

TAP Gibbs Free Energy, Belief Propagation and Sparsity

Lehel Csató, Manfred Opper, Ole Winther

The adaptive TAP Gibbs free energy for a general densely connected probabilistic model with quadratic interactions and arbritary single site constraints is derived. We show how a speci■c sequential minimization of the free energy leads to a generalization of Minka's expectation propa- gation. Lastly, we derive a sparse representation version of the sequential algorithm. The usefulness of the approach is demonstrated on classi■ca- tion and density estimation with Gaussian processes and on an indepen- dent component analysis problem.

****

Quantizing Density Estimators

Peter Meinicke, Helge Ritter

We suggest a nonparametric framework for unsupervised learning of projection models in terms of density estimation on quantized sample spaces. The objective is not to optimally reconstruct the data but in- stead the quantizer is chosen to optimally reconstruct the density of the data. For the resulting quantizing density estimator (QDE) we present a general method for parameter estimation and model selection. We show how projection sets which correspond to traditional unsupervised meth- ods like vector quantization or PCA appear in the new framework. For a principal component quantizer we present results on synthetic and real- world data, which show that the QDE can improve the generalization of the kernel density estimator although its estimate is based on signi■cantly lower-dimensional projection indices of the data.

****

The Concave-Convex Procedure (CCCP)

Alan L. Yuille, Anand Rangarajan

We introduce the Concave-Convex procedure (CCCP) which con(cid:173)structs discrete time iterative dynamical systems which are guar(cid:173)anteed to monotonically decrease global optimization/energy func(cid:173)tions. It can be applied to (almost) any optimization problem and many existing algorithms can be interpreted in terms of CCCP. In particular, we prove relationships to some applications of Legendre transform techniques. We then illustrate CCCP by applications to Potts models, linear assignment, EM algorithms, and Generalized Iterative Scaling (GIS). CCCP can be used both as a new way to understand existing optimization algorithms and as a procedure for generating new algorithms.

****

Active Learning in the Drug Discovery Process

Manfred K. K. Warmuth, Gunnar Rätsch, Michael Mathieson, Jun Liao, Christian Lemmen

We investigate the following data mining problem from Computational Chemistry: From a large data set of compounds, ■nd those that bind to a target molecule in as few iterations of biological testing as possible. In each iteration a comparatively small batch of compounds is screened for binding to the target. We apply active learning techniques for selecting the successive batches. One selection strategy picks unlabeled examples closest to the maximum margin hyperplane. Another produces many weight vectors by running perceptrons over multiple permutations of the data. Each weight vector prediction and we pick the unlabeled examples for which votes with its the prediction is most evenly split between . For a third selec- tion strategy note that each unlabeled example bisects the version space of consistent weight vectors. We estimate the volume on both sides of the split by bouncing a billiard through the version space and select un- labeled examples that cause the most even split of the version space. We demonstrate that on two data sets provided by DuPont Pharmaceu- ticals that all three selection strategies perform comparably well and are much better than selecting random batches

for testing.
****
Intransitive Likelihood-Ratio Classifiers
Jeff Bilmes, Gang Ji, Marina Meila
In this work, we introduce an information-theoretic based correction term to the likelihood ratio classi■cation method for multiple classes. Under certain conditions, the term is suf■cient for optimally correcting the dif- ference between the true and estimated likelihood ratio, and we analyze this in the Gaussian case. We ■nd that the new correction term signif- icantly improves the classi■cation results when tested on medium vo- cabulary speech recognition tasks. Moreover, the addition of this term makes the class comparisons analogous to an intransitive game and we therefore use several tournament-like strategies to deal with this issue. We ■nd that further small improvements are obtained by using an appro- priate tournament. Lastly, we ■nd that intransitivity appears to be a good measure of classi■cation con■dence.
****
Transform-invariant Image Decomposition with Similarity Templates
Chris Stauffer, Erik Miller, Kinh Tieu
Recent work has shown impressive transform-invariant modeling and clustering for sets of images of objects with similar appearance. We seek to expand these capabilities to sets of images of an object class that show considerable variation across individual instances (e.g. pedestrian images) using a representation based on pixel-wise similarities, similarity templates. Because of its invariance to the colors of particular components of an object, this representation en- ables detection of instances of an object class and enables alignment of those instances. Further, this model implicitly represents the re- gions of color regularity in the class-speci(cid:12)c image set enabling a decomposition of that object class into component regions.
****
Activity Driven Adaptive Stochastic Resonance
Gregor Wenning, Klaus Obermayer
Cortical neurons might be considered as threshold elements inte(cid:173) grating in parallel many excitatory and inhibitory inputs. Due to the apparent variability of cortical spike trains this yields a strongly fluctuating membrane potential, such that threshold crossings are highly irregular. Here we study how a neuron could maximize its sensitivity w.r.t. a relatively small subset of excitatory input. Weak signals embedded in fluctuations is the natural realm of stochastic resonance. The neuron's response is described in a hazard-function approximation applied to an Ornstein-Uhlenbeck process. We an(cid:173) alytically derive an optimality criterium and give a learning rule for the adjustment of the membrane fluctuations, such that the sensitivity is maximal exploiting stochastic resonance. We show that adaptation depends only on quantities that could easily be estimated locally (in space and time) by the neuron. The main results are compared with simulations of a biophysically more re(cid:173) alistic neuron model.
****
Reinforcement Learning with Long Short-Term Memory
Bram Bakker
This paper presents reinforcement learning with a Long Short(cid:173) Term Memory recurrent neural network: RL-LSTM. Model-free RL-LSTM using Advantage(,x) learning and directed exploration can solve non-Markovian tasks with long-term dependencies be(cid:173) tween relevant events. This is demonstrated in a T-maze task, as well as in a difficult variation of the pole balancing task.
****
A Dynamic HMM for On-line Segmentation of Sequential Data
Jens Kohlmorgen, Steven Lemm
We propose a novel method for the analysis of sequential data that exhibits an inherent mode switching. In particular, the data might be a non-stationary time series from a dynamical system that switches between multiple operating modes. Unlike other ap(cid:173) proaches, our method proces

ses the data incrementally and without any training of internal parameters. We use an HMM with a dy(cid:173) namically changing number of states and a n on-line variant of the Viterbi algorithm that performs an unsupervised segmen tation and classification of the data on-the-fly, i.e. the method is able to pro(cid:173) cess incoming data in real-time. The main idea of the approach is to track and segment changes of the probability density of the data in a sliding window on the incoming data stream. The usefulness of the algorithm is demonstrated by an application to a switching dynamical system.
****

The Intelligent surfer: Probabilistic Combination of Link and Content Informatio n in PageRank
Matthew Richardson, Pedro Domingos
The PageRank algorithm, used in the Google search engine, greatly improves the results of Web search by taking into account the link structure of the Web. PageRank assigns to a page a score propor- tional to the number of times a random surfer would visit that page, if it surfed indefinitely from pag e to page, following all outlinks from a page with equal probability. We propose to improve Page- Rank by using a more intelligent surfer, one that is guided by a probabilistic model of the relevance of a page t o a query. Efficient execution of our algorithm at query time is made possible by pre- computing at crawl time (and thus once for all queri es) the neces- sary terms. Experiments on two large subsets of the We b indicate that our algorithm significantly outperforms PageRank in the (hu- man-rated) quality of the pages returned, while remaining efficient en ough to be used in today’s large search engines.
****

Incremental A*
S. Koenig, M. Likhachev
Incremental search techniques ■nd optimal solutions to series of similar search tasks much faster than is possible by solving each search task from scratch. Whi le researchers have developed incremental versions of uninformed search methods, we develop an incremental version of A. The ■rst search of Lifelong Planning A is the same as that of A* but all subsequent searches are much faster because it reuses those parts of the previous search tree that are identical to the new se arch tree. We then present experimental results that demonstrate the advantages of Lifelong Planning A* for simple route planning tasks.
****

Grouping and dimensionality reduction by locally linear embedding
Marzia Polito, Pietro Perona
(LLE)
****

Partially labeled classification with Markov random walks
Martin Szummer, Tommi Jaakkola
To classify a large number of unlabeled examples we combine a lim- ited number o f labeled examples with a Markov random walk represen- tation over the unlabeled examples. The random walk representation ex- ploits any low dimensional structu re in the data in a robust, probabilistic manner. We develop and compare several estimation criteria/algorithms suited to this representation. This includes in particular multi-way clas- si■cation with an average margin criterion which perm its a closed form solution. The time scale of the random walk regularizes the re presenta- tion and can be set through a margin-based criterion favoring unambigu - ous classi■cation. We also extend this basic regularization by adapting time s cales for individual examples. We demonstrate the approach on synthetic examples and on text classi■cation problems.
****

Dynamic Time-Alignment Kernel in Support Vector Machine
Hiroshi Shimodaira, Ken-ichi Noma, Mitsuru Nakai, Shigeki Sagayama
A new class of Support Vector Machine (SVM) that is applica- ble to sequential-p attern recognition such as speech recognition is developed by incorporating an i dea of non-linear time alignment into the kernel function. Since the time-alignm

ent operation of sequential pattern is embedded in the new kernel function, stan - dard SVM training and classification algorithms can be employed without furthe r modifications. The proposed SVM (DTAK-SVM) is evaluated in speaker-dependent s peech recognition experiments of hand-segmented phoneme recognition. Preliminary experimen- tal results show comparable recognition performance with hidden Mark ov models (HMMs). 1 Introduction

****

Associative memory in realistic neuronal networks

Peter Latham

Almost two decades ago, Hopfield [1] showed that networks of highly re duced model neurons can exhibit multiple attracting fixed points, thus providin g a substrate for associative memory. It is still not clear, however, whether realistic neuronal networks can support multiple attractors. The main difficu lty is that neuronal networks in vivo exhibit a stable background state at low firing rate, typ(cid:173) ically a few Hz. Embedding attractor is easy; doing so without destabilizing the background is not. Previ ous work [2, 3] focused on the sparse coding limit, in which a vanishingly small number of neurons are involved in any memory. Here we investigate t he case in which the number of neurons involved in a memory scales with t he number of neurons in the network. In contrast to the sparse coding limit, we find that multiple attractors can co-exist robustly with a stable background state. Mean field theory is used to under(cid:173) stand how the b ehavior of the network scales with its parameters, and simulations with analog neurons are presented.

****

A General Greedy Approximation Algorithm with Applications

T. Zhang

Greedy approximation algorithms have been frequently used to obtain sparse solut ions to learning problems. In this paper, we present a general greedy algorithm for solving a class of convex optimization problems. We derive a bound on the ra te of approximation for this algorithm, and show that our algorithm includes a n umber of earlier studies as special cases.

****

Reducing multiclass to binary by coupling probability estimates

B. Zadrozny

This paper presents a method for obtaining class membership probability esti- ma tes for multiclass classi■cation problems by coupling the probability estimates produced by binary classi■ers. This is an extension for arbitrary code matrices of a method due to Hastie and Tibshirani for pairwise coupling of probability es timates. Experimental results with Boosted Naive Bayes show that our method prod uces calibrated class membership probability estimates, while having similar cla ssi■cation accuracy as loss-based decoding, a method for obtaining the most like ly class that does not generate probability estimates.

****

Linear-time inference in Hierarchical HMMs

Kevin P. Murphy, Mark Paskin

The hierarchical hidden Markov model (HHMM) is a generalization of the hidden Ma rkov model (HMM) that models sequences with structure at many length/time scales [FST98]. Unfortunately, the original infer- is ence algorithm is rather complic ated, and takes the length of the sequence, making it impractical for many domai ns. In this paper, we show how HHMMs are a special kind of dynamic Bayesian netw ork (DBN), and thereby derive a much simpler inference algorithm, which only tak es time. Furthermore, by drawing the connection between HHMMs and DBNs, we enabl e the application of many stan- dard approximation techniques to further speed u p inference.

****

Sequential Noise Compensation by Sequential Monte Carlo Method

K. Yao, S. Nakamura

We present a sequential Monte Carlo method applied to additive noise compensatio n for robust speech recognition in time-varying noise. The method generates a se

t of samples according to the prior distribution given by clean speech models an d noise prior evolved from previous estimation. An explicit model representing n oise ef- fects on speech features is used, so that an extended Kalman filter is constructed for each sample, generating the updated continuous state estimate as the estimation of the noise parameter, and predic- tion likelihood for weightin g each sample. Minimum mean square error (MMSE) inference of the time-varying no ise parameter is car- ried out over these samples by fusion the estimation of sa mples ac- cording to their weights. A residual resampling selection step and a M etropolis-Hastings smoothing step are used to improve calcula- tion e#ciency. Ex periments were conducted on speech recognition in simulated non-stationary noise s, where noise power changed ar- tificially, and highly non-stationary Machinegu n noise. In all the experiments carried out, we observed that the method can hav e sig- nificant recognition performance improvement, over that achieved by noise compensation with stationary noise assumption. 1 Introduction

****

Direct value-approximation for factored MDPs

Dale Schuurmans, Relu Patrascu

We present a simple approach for computing reasonable policies for factored Mark ov decision processes (MDPs), when the opti(cid:173) mal value function can be a pproximated by a compact linear form. Our method is based on solving a single li near program that ap(cid:173) proximates the best linear fit to the optimal valu e function. By applying an efficient constraint generation procedure we obtain a n iterative solution method that tackles concise linear programs. This direct li near programming approach experimentally yields a signif(cid:173) icant reductio n in computation time over approximate value- and policy-iteration methods (some times reducing several hours to a few seconds). However, the quality of the solu tions produced by linear programming is weaker-usually about twice the approxi(c id:173) mation error for the same approximating class. Nevertheless, the speed a dvantage allows one to use larger approximation classes to achieve similar error in reasonable time.

****

Perceptual Metamers in Stereoscopic Vision

B. Backus

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.

****

Scaling Laws and Local Minima in Hebbian ICA

Magnus Rattray, Gleb Basalyga

We study the dynamics of a Hebbian ICA algorithm extracting a sin- gle non-Gauss ian component from a high-dimensional Gaussian back- ground. For both on-line an d batch learning we ■nd that a surprisingly large number of examples are require d to avoid trapping in a sub-optimal state close to the initial conditions. To e xtract a skewed signal at least examples are required for -dimensional data an d

****

Online Learning with Kernels

Jyrki Kivinen, Alex Smola, Robert C. Williamson

We consider online learning in a Reproducing Kernel Hilbert Space. Our method is computationally ef■cient and leads to simple algorithms. In particular we deriv e update equations for classi■cation, regression, and novelty detection. The inc lusion of the -trick allows us to give a robust parameterization. Moreover, unl ike in batch learning where the -trick only applies to the -insensitive loss fu nction we are able to derive gen- eral trimmed-mean types of estimators such as for Huber’s robust loss.

****

3 state neurons for contextual processing

Ádám Kepecs, S. Raghavachari

Neurons receive excitatory inputs via both fast AMPA and slow NMDA ty

pe receptors. We find that neurons receiving input via NMDA receptors can have two stable membrane states which are input dependent. Action potentials can only be initiated from the higher voltage state. Similar observations have been made in sev(cid:173) eral brain areas which might be explained by our model. The in(cid:173) teractions between the two kinds of inputs lead us to suggest that some neurons may operate in 3 states: disabled, enabled and fir(cid:173) ing. Such enabled, but non-firing modes can be used to introduce context-dependent processing in neural networks. We provide a simple example and discuss possible implications for neuronal pro(cid:173) cessing and response variability.

****

Convolution Kernels for Natural Language

Michael Collins, Nigel Duffy

We describe the application of kernel methods to Natural Language Pro- cessing (NLP) problems. In many NLP tasks the objects being modeled are strings, trees, graphs or other discrete structures which require some mechanism to convert them into feature vectors. We describe kernels for various natural language structures, allowing rich, high dimensional rep- resentations of these structures. We show how a kernel over trees can be applied to parsing using the voted perceptron algorithm, and we give experimental results on the ATIS corpus of parse trees.

****

On the Convergence of Leveraging

Gunnar Rätsch, Sebastian Mika, Manfred K. K. Warmuth

We give an unified convergence analysis of ensemble learning meth- ods including e.g. AdaBoost, Logistic Regression and the Least-Square- Boost algorithm for regression. These methods have in common that they iteratively call a base learning algorithm which returns hypotheses that are then linearly combined. We show that these methods are related to the Gauss-Southwell method known from numerical optimization and state non-asymptotical convergence results for all these methods. Our analysis includes ` 1 -norm regularized cost functions leading to a clean and general way to regularize ensemble learning. 1 Introduction

****

A Quantitative Model of Counterfactual Reasoning

Daniel Yarlett, Michael Ramscar

In this paper we explore two quantitative approaches to the modelling of counterfactual reasoning – a linear and a noisy-OR model – based on in- formation contained in conceptual dependency networks. Empirical data is acquired in a study and the ■t of the models compared to it. We con- clude by considering the appropriateness of non-parametric approaches to counterfactual reasoning, and examining the prospects for other para- metric approaches in the future.

****

A Neural Oscillator Model of Auditory Selective Attention

Stuart Wrigley, Guy Brown

A model of auditory grouping is described in which auditory attention plays a key role. The model is based upon an oscillatory correlation framework, in which neural oscillators representing a single perceptual stream are synchronised, and are desynchronised from oscillators representing other streams. The model suggests a mechanism by which attention can be directed to the high or low tones in a repeating sequence of tones with alternating frequencies. In addition, it simulates the perceptual segregation of a mistuned harmonic from a complex tone.

****

Self-regulation Mechanism of Temporally Asymmetric Hebbian Plasticity

N. Matsumoto, M. Okada

Recent biological experimental (cid:12)ndings have shown that the synap- tic plasticity depends on the relative timing of the pre- and post- synaptic spikes which determines whether Long Term Potentiation (LTP) occurs or Long Term Depression (LTD) does. The synaptic plasticity has been called \Temporally Asymmetric Hebbian plas- ticity (TAH)". Many authors have numerically shown that spatio- temporal patterns can be stored in neural networks. However, the mathematical mechanism for storage of the spatio-temporal pat- terns is still unknown, especially th

e e(cid:11)ects of LTD. In this paper, we employ a simple neural network model a nd show that inter- ference of LTP and LTD disappears in a sparse coding scheme. On the other hand, it is known that the covariance learning is in- dispensable for storing sparse patterns. We also show that TAH qualitatively has the same e(cid:11)ect as the covariance learning when spatio-temporal patterns are embedded in the network.

****

An Efficient, Exact Algorithm for Solving Tree-Structured Graphical Games

Michael Littman, Michael Kearns, Satinder Singh

We describe a new algorithm for computing a Nash equilibrium in graphical games, a compact representation for multi-agent systems that we introduced in previous work. The algorithm is the first to compute equilibria bot h efficiently and exactly for a non-trivial class of graphical games.

****

Causal Categorization with Bayes Nets

Bob Rehder

A theory of categorization is presented in which knowledge of causal relationships between category features is represented as a Bayesian ne twork. Referred to as causal-model theory, this theory predicts that o bjects are classified as category members to the extent they are likel y to have been produced by a categorys causal model. On this view, p eople have models of the world that lead them to expect a certain d istribution of features in category members (e.g., correlations between feature pairs that are directly connected by causal relationships), an d consider exemplars good category members when they manifest those ex pectations. These expectations include sensitivity to higher-order featur e interactions that emerge from the asymmetries inherent in causal relationsh ips.

****

KLD-Sampling: Adaptive Particle Filters

Dieter Fox

Over the last years, particle ■lters have been applied with great success to a v ariety of state estimation problems. We present a statistical approach to increa sing the ef■ciency of particle ■lters by adapting the size of sample sets on-the -■y. The key idea of the KLD-sampling method is to bound the approximation error introduced by the sample-based representation of the particle ■lter. The name K LD-sampling is due to the fact that we measure the approximation error by the Ku llback-Leibler distance. Our adaptation approach chooses a small number of sampl es if the density is focused on a small part of the state space, and it chooses a large number of samples if the state uncertainty is high. Both the implementat ion and computation overhead of this approach are small. Extensive experiments u sing mobile robot localization as a test application show that our approach yiel ds drastic improvements over particle ■lters with ■xed sample set sizes and over a previously introduced adaptation technique.

****

Unsupervised Learning of Human Motion Models

Yang Song, Luis Goncalves, Pietro Perona

This paper presents an unsupervised learning algorithm that can derive the proba bilistic dependence structure of parts of an object (a moving hu- man body in ou r examples) automatically from unlabeled data. The dis- tinguished part of this work is that it is based on unlabeled data, i.e., the training features include both useful foreground parts and background clutter and the correspondence betwe en the parts and detected features are unknown. We use decomposable triangulated graphs to depict the probabilistic independence of parts, but the unsupervised technique is not limited to this type of graph. In the new approach, labeling of the data (part assignments) is taken as hidden variables and the EM algo- rithm is applied. A greedy algorithm is developed to select parts and to search for t he optimal structure based on the differential entropy of these variables. The s uccess of our algorithm is demonstrated by applying it to generate models of hum an motion automatically from unlabeled real image sequences.

****
Orientational and Geometric Determinants of Place and Head-direction
Neil Burgess, Tom Hartley
We present a model of the firing of place and head-direction cells in rat hippoc ampus. The model can predict the response of individual cells and populations to parametric manipulations of both geomet(cid:173) ric (e.g. O'Keefe & Burgess, 1 996) and orientational (Fenton et aI., 2000a) cues, extending a previous geometr ic model (Hartley et al., 2000). It provides a functional description of how the se cells' spatial responses are derived from the rat's environment and makes eas ily testable quantitative predictions. Consideration of the phenomenon of remapp ing (Muller & Kubie, 1987; Bostock et aI., 1991) indicates that the model may al so be consistent with non(cid:173) parametric changes in firing, and provides co nstraints for its future development.
****
Covariance Kernels from Bayesian Generative Models
Matthias Seeger
We propose the framework of mutual information kernels for learning co variance kernels, as used in Support Vector machines and Gaussian proc ess classifiers, from unlabeled task data using Bayesian techniques. We describe an implementation of this frame(cid:173) work which uses variational Bayesian mixtures of factor analyzers in order to attack classification proble ms in high-dimensional spaces where labeled data is sparse, but unlabeled data is abundant.
****
Modeling Temporal Structure in Classical Conditioning
Aaron C. Courville, David Touretzky
The Temporal Coding Hypothesis of Miller and colleagues [7] sug(cid:173) gest s that animals integrate related temporal patterns of stimuli into sing le memory representations. We formalize this concept using quasi-Bayes estimation to update the parameters of a con(cid:173) strained hidden Mark ov model. This approach allows us to account for some surprising temporal eff ects in the second order condition(cid:173) ing experiments of Miller et al . [1 , 2, 3], which other models are unable to explain.
****
Modeling the Modulatory Effect of Attention on Human Spatial Vision
Laurent Itti, Jochen Braun, Christof Koch
We present new simulation results, in which a computational model of interactin g visual neurons simultaneously predicts the modula(cid:173) tion of spatial vis ion thresholds by focal visual attention, for five dual-task human psychophysic s experiments. This new study com(cid:173) plements our previous findings that a ttention activates a winner(cid:173) take-all competition among early visual neu rons within one cortical hypercolumn. This "intensified competition" hypothesis assumed that attention equally affects all neurons, and yielded two single(cid :173) unit predictions: an increase in gain and a sharpening of tuning with att ention. While both effects have been separately observed in electrophysiology, no single-unit study has yet shown them si(cid:173) multaneously. Hence, we here explore whether our model could still predict our data if attention might only modulate neuronal gain, but do so non-uniformly across neurons and tasks. Spec ifically, we investigate whether modulating the gain of only the neurons that are loudest, best-tuned, or most informative about the stimulus, or of all neur ons equally but in a task-dependent manner, may ac(cid:173) count for the data. We find that none of these hypotheses yields predictions as plausible as the in tensified competition hypothesis, hence providing additional support for our or iginal findings.
****
Approximate Dynamic Programming via Linear Programming
Daniela Farias, Benjamin Roy
The curse of dimensionality gives rise to prohibitive computational requirement s that render infeasible the exact solution of large- scale stochastic control problems. We study an efficient method based on linear programming fo

r approximating solutions to such prob(cid:173) lems. The approach "fits " a linear combination of pre- selected basis functions to the dynamic programming cost- to- go function. We develop bounds on the approximation err or and present experi(cid:173) mental results in the domain of queueing network control, providing empirical support for the methodology.

****

A New Discriminative Kernel From Probabilistic Models

Koji Tsuda, Motoaki Kawanabe, Gunnar Rätsch, Sören Sonnenburg, Klaus-Robert Müll er

Recently, Jaakkola and Haussler proposed a method for construct(cid:173) ing kernel functions from probabilistic models. Their so called "Fisher ker nel" has been combined with discriminative classifiers such as SVM and applied successfully in e.g. DNA and protein analysis. Whereas the F isher kernel (FK) is calculated from the marginal log-likelihood, we p ropose the TOP kernel derived from Tangent vectors Of Posterior log-od ds. Furthermore we develop a theoretical framework on feature extractor s from probabilistic models and use it for analyzing FK and TOP. In ex periments our new discriminative TOP kernel compares favorably to the F isher kernel.

****

Kernel Feature Spaces and Nonlinear Blind Souce Separation

Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, Klaus-Robert Müller

In kernel based learning the data is mapped to a kernel feature space of a dimen sion that corresponds to the number of training data points. In practice, howeve r, the data forms a smaller submanifold in feature space, a fact that has been u sed e.g. by reduced set techniques for SVMs. We propose a new mathematical const ruction that permits to adapt to the in- trinsic dimension and to ■nd an orthono rmal basis of this submanifold. In doing so, computations get much simpler and m ore important our theoretical framework allows to derive elegant kernelized blin d source separation (BSS) algorithms for arbitrary invertible nonlinear mixings. Experiments demonstrate the good performance and high computational ef■ciency o f our kTDSEP algorithm for the problem of nonlinear BSS.

****

The Unified Propagation and Scaling Algorithm

Yee Teh, Max Welling

In this paper we will show that a restricted class of constrained mini- mum dive rgence problems, named generalized inference problems, can be solved by approxim ating the KL divergence with a Bethe free energy. The algorithm we derive is clo sely related to both loopy belief propaga- tion and iterative scaling. This uni■ ed propagation and scaling algorithm reduces to a convergent alternative to loop y belief propagation when no constraints are present. Experiments show the viabi lity of our algorithm.

****

Receptive field structure of flow detectors for heading perception

J. Beintema, M. Lappe, Alexander Berg

Observer translation relative to the world creates image flow that expands fr om the observer's direction of translation (heading) from which the observer can recover heading direction. Yet, the image flow is often more compl ex, depending on rotation of the eye, scene layout and translation velocity . A number of models [1-4] have been proposed on how the human visua l system extracts heading from flow in a neurophysiologic ally plausible way. These models represent heading by a set of neurons that respond to l arge image flow patterns and receive input from motion sensed at different im(cid:173) age locations. We analysed these models to determine the exact receptive field of these heading detectors. We find most models predi ct that, contrary to widespread believe, the contribut ing mo(cid:173) tion sen sors have a preferred motion directed circularly rather than radially around th e detector's preferred heading. Moreover, the re(cid:173) sults suggest to look for more refined structure within the circular flow, such as bi-circularity or local motion-opponency.

****
Bayesian morphometry of hippocampal cells suggests same-cell somatodendritic repulsion
Giorgio Ascoli, Alexei Samsonovich
Visual inspection of neurons suggests that dendritic orientation may be determined both by internal constraints (e.g. membrane tension) and by external vector fields (e.g. neurotrophic gradients). For example, basal dendrites of pyramidal cells appear nicely fan-out. This regular orientation is hard to justify completely with a general tendency to grow straight, given the zigzags observed experimentally. Instead, dendrites could (A) favor a fixed ("external") direction, or (B) repel from their own soma. To investigate these possibilities quantitatively, reconstructed hippocampal cells were subjected to Bayesian analysis. The statistical model combined linearly factors A and B, as well as the tendency to grow straight. For all morphological classes, B was found to be significantly positive and consistently greater than A. In addition, when dendrites were artificially re-oriented according to this model, the resulting structures closely resembled real morphologies. These results suggest that somatodendritic repulsion may play a role in determining dendritic orientation. Since hippocampal cells are very densely packed and their dendritic trees highly overlap, the repulsion must be cell- specific. We discuss possible mechanisms underlying such specificity.
****
Entropy and Inference, Revisited
Ilya Nemenman, F. Shafee, William Bialek
We study properties of popular near–uniform (Dirichlet) priors for learn- ing undersampled probability distributions on discrete nonmetric spaces and show that they lead to disastrous results. However, an Occam–style phase space argument expands the priors into their in■nite mixture and resolves most of the observed problems. This leads to a surprisingly good estimator of entropies of discrete distributions.
****
Audio-Visual Sound Separation Via Hidden Markov Models
John Hershey, Michael Casey
It is well known that under noisy conditions we can hear speech much more clearly when we read the speaker's lips. This sug(cid:173) gests the utility of audio-visual information for the task of speech enhancement. We propose a method to exploit audio-visual cues to enable speech separation under non-stationary noise and with a single microphone. We revise and extend HMM-based speech enhancement techniques, in which signal and noise models are fac(cid:173) tori ally combined, to incorporate visual lip information and em(cid:173) ploy novel signal HMMs in which the dynamics of narrow-band and wide band components are factorial. We avoid the combina(cid:173) torial explosion in the factorial model by using a simple approxi(cid:173) mate inference technique to quickly estimate the clean signals in a mixture. We present a preliminary evaluation of this approach using a small-vocabulary audio-visual database, showing promising improvements in machine intelligibility for speech enhanced using audio and visual information.
****
Spectral Relaxation for K-means Clustering
Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, Horst Simon
The popular K-means clustering partitions a data set by minimiz(cid:173) ing a sum-of-squares cost function. A coordinate descend method is then used to find local minima. In this paper we show that the minimization can be reformulated as a trace maximization problem associated with the Gram matrix of the data vectors. Furthermore, we show that a relaxed version of the trace maximization problem possesses global optimal solutions which can be obtained by com(cid:173) puting a partial eigendecomposition of the Gram matrix, and the cluster assignment for each data vectors can be found by comput(cid:173) ing a pivoted QR decomposition of the eigenvector matrix. As a by-p

roduct we also derive a lower bound for the minimum of the sum-of-squa
res cost function.
****

Gaussian Process Regression with Mismatched Models
Peter Sollich
Learning curves for Gaussian process regression are well understood when the 's
tudent' model happens to match the 'teacher' (true data generation process). I
derive approximations to the learning curves for the more generic case of mis
matched models, and find very rich behaviour: For large input space dimension
ality, where the results become exact, there are universal (student-indep
endent) plateaux in the learning curve, with transitions in between that can
exhibit arbitrarily many over-fitting maxima; over-fitting can occur ev
en if the student estimates the teacher noise level correctly. In lower di
mensions, plateaux also appear, and the learning curve remains dependent on
the mismatch between student and teacher even in the asymptotic limit of
a large number of training examples. Learn(cid:173) ing with excessively stron
g smoothness assumptions can be partic(cid:173) ularly dangerous: For example
, a student with a standard radial basis function covariance function will
learn a rougher teacher func(cid:173) tion only logarithmically slowly. All
predictions are confirmed by simulations.
****

Kernel Machines and Boolean Functions
Adam Kowalczyk, Alex Smola, Robert C. Williamson
We give results about the learnability and required complexity of logical formul
ae to solve classi■cation problems. These results are obtained by linking propos
itional logic with kernel machines. In particular we show that decision trees an
d disjunctive normal forms (DNF) can be repre- sented by the help of a special k
ernel, linking regularized risk to separa- tion margin. Subsequently we derive a
number of lower bounds on the required complexity of logic formulae using prope
rties of algorithms for generation of linear estimators, such as perceptron and
maximal percep- tron learning.
****

Efficient Resources Allocation for Markov Decision Processes
Rémi Munos
It is desirable that a complex decision-making problem in an uncer(cid:173) tain
world be adequately modeled by a Markov Decision Process (MDP) whose structural
representation is adaptively designed by a parsimonious resources allocation pr
ocess. Resources include time and cost of exploration, amount of memory and comp
utational time allowed for the policy or value function representation. Concerne
d about making the best use of the available resources, we address the problem o
f efficiently estimating where adding extra resources is highly needed in order
to improve the expected performance of the resulting policy. Possible applicatio
n in reinforcement learning (RL) , when real-world exploration is highly costly,
concerns the de(cid:173) tection of those areas of the state-space that need pr
imarily to be explored in order to improve the policy. Another application con(c
id:173) cerns approximation of continuous state-space stochastic control problem
s using adaptive discretization techniques for which highly efficient grid point
s allocation is mandatory to survive high dimen(cid:173) sionality. Maybe surpri
singly these two problems can be formu(cid:173) lated under a common framework:
for a given resource allocation, which defines a belief state over possible MDPs
, find where adding new resources (thus decreasing the uncertainty of some param
e(cid:173) ters -transition probabilities or rewards) will most likely increase
the expected performance of the new policy. To do so, we use sam(cid:173) pling
techniques for estimating the contribution of each parameter's probability distr
ibution function (Pdf) to the expected loss of us(cid:173) ing an approximate po
licy (such as the optimal policy of the most probable MDP) instead of the true (
but unknown) policy.
****

Information Geometrical Framework for Analyzing Belief Propagation Decoder
Shiro Ikeda, Toshiyuki Tanaka, Shun-ichi Amari

The mystery of belief propagation (BP) decoder, especially of the turbo decoding , is studied from information geometrical viewpoint. The loopy belief network (B N) of turbo codes makes it dif■cult to obtain the true “belief” by BP, and the c haracteristics of the algorithm and its equilib- rium are not clearly understood . Our study gives an intuitive understand- ing of the mechanism, and a new frame work for the analysis. Based on the framework, we reveal basic properties of the turbo decoding.

****

Speech Recognition using SVMs

N. Smith, Mark Gales

An important issue in applying SVMs to speech recognition is the ability to classify variable length sequences. This paper presents extensions to a standard scheme for handling this variable length data, the Fisher score. A more useful mapping is introduced based on the likelihood-ratio. The score -space defined by this mapping avoids some limitations of the Fisher score. Class-conditional gen(cid:173) erative models are directly incorporated int o the definition of the score-space. The mapping, and appropriate normalisat ion schemes, are evaluated on a speaker-independent isolated letter task whe re the new mapping outperforms both the Fisher score and HMMs trained to maximise likelihood.

****

Learning Discriminative Feature Transforms to Low Dimensions in Low Dimentions

Kari Torkkola

The marriage of Renyi entropy with Parzen density estimation has been shown to b e a viable tool in learning discriminative feature transforms. However, it suffe rs from computational complexity proportional to the square of the number of sam ples in the training data. This sets a practical limit to using large databases. We suggest immediate divorce of the two methods and remarriage of Renyi entropy with a semi-parametric density estimation method, such as a Gaussian Mixture Mo dels (GMM). This al- lows all of the computation to take place in the low dimens ional target space, and it reduces computational complexity proportional to squa re of the number of components in the mixtures. Furthermore, a conve- nient exte nsion to Hidden Markov Models as commonly used in speech recognition becomes pos sible.

****

Information-Geometrical Significance of Sparsity in Gallager Codes

Toshiyuki Tanaka, Shiro Ikeda, Shun-ichi Amari

We report a result of perturbation analysis on decoding error of the belief prop agation decoder for Gallager codes. The analysis is based on infor- mation geome try, and it shows that the principal term of decoding error at equilibrium comes from the m-embedding curvature of the log-linear submanifold spanned by the est imated pseudoposteriors, one for the full marginal, and K for partial posteriors , each of which takes a single check into account, where K is the number of chec ks in the Gallager code. It is then shown that the principal error term vanishes when the parity-check matrix of the code is so sparse that there are no two col umns with overlap greater than 1.

****

Adaptive Sparseness Using Jeffreys Prior

Mário Figueiredo

In this paper we introduce a new sparseness inducing prior which does not involv e any (hy- per)parameters that need to be adjusted or estimated. Although other applications are possi- ble, we focus here on supervised learning problems: regr ession and classi■cation. Experi- ments with several publicly available benchmar k data sets show that the proposed approach yields state-of-the-art performance. In particular, our method outperforms support vector machines and performs comp etitively with the best alternative techniques, both in terms of error rates and sparseness, although it involves no tuning or adjusting of sparseness- controll ing hyper-parameters.

****

An Efficient Clustering Algorithm Using Stochastic Association Model and Its Imp

lementation Using Nanostructures
Takashi Morie, Tomohiro Matsuura, Makoto Nagata, Atsushi Iwata
This paper describes a clustering algorithm for vector quantizers using a “stoch astic association model”. It offers a new simple and powerful soft- max adaptati on rule. The adaptation process is the same as the on-line K-means clustering me thod except for adding random ■uctuation in the distortion error evaluation proc ess. Simulation results demonstrate that the new algorithm can achieve ef■cient adaptation as high as the “neural gas” algorithm, which is reported as one of th e most ef■cient clustering methods. It is a key to add uncorrelated random ■uctu ation in the simi- larity evaluation process for each reference vector. For hard ware imple- mentation of this process, we propose a nanostructure, whose operati on is described by a single-electron circuit. It positively uses ■uctuation in q uantum mechanical tunneling processes.
****

Grammar Transfer in a Second Order Recurrent Neural Network
Michiro Negishi, Stephen Hanson
It has been known that people, after being exposed to sentences gener ated by an artificial grammar, acquire implicit grammatical knowledge an d are able to transfer the knowledge to inputs that are generated by a modif ied grammar. We show that a second order recurrent neural network is abl e to transfer grammatical knowledge from one language (generated by a Finite St ate Machine) to another language which differ both in vocabularies and syntax. Representa(cid:173) tion of the grammatical knowledge in the network is analyz ed using linear discriminant analysis.
****

Product Analysis: Learning to Model Observations as Products of Hidden Variables
Brendan J. Frey, Anitha Kannan, Nebojsa Jojic
Factor analysis and principal components analysis can be used to model li near relationships between observed variables and linearly map high-dimension al data to a lower-dimensional hidden space. In factor analysis, the o bservations are modeled as a linear com(cid:173) bination of normally di stributed hidden variables. We describe a nonlinear generalization of fac tor analysis, called "product analy(cid:173) sis", that models the observ ed variables as a linear combination of products of normally distributed hidden variables. Just as fac(cid:173) tor analysis can be viewed as unsupervised linear regression on unobserved, normally distributed hidde n variables, product anal(cid:173) ysis can be viewed as unsupervised l inear regression on products of unobserved, normally distributed hidden variables. The map(cid:173) ping between the data and the hidden space is nonlinear, so we use an approximate variational technique for infer ence and learn(cid:173) ing. Since product analysis is a generalization of factor analysis, product analysis always finds a higher data likeliho od than factor analysis. We give results on pattern recognition and illumina tion(cid:173) invariant image clustering.
****

Matching Free Trees with Replicator Equations
Marcello Pelillo
Motivated by our recent work on rooted tree matching, in this paper we provide a solution to the problem of matching two free (i.e., unrooted) trees by construc ting an association graph whose maximal cliques are in one-to-one correspondence with maximal common subtrees. We then solve the problem using simple replicator dynamics from evolutionary game theory. Experiments on hundreds of uniformly ra ndom trees are presented. The results are impressive: despite the inherent inabi lity of these simple dynamics to escape from local optima, they always returned a globally optimal solution.
****

Multiplicative Updates for Classification by Mixture Models
Lawrence Saul, Daniel Lee
We investigate a learning algorithm for the classi■cation of nonnegative data by mixture models. Multiplicative update rules are derived that directly optimize

the performance of these models as classi■ers. The update rules have a simple cl osed form and an intuitive appeal. Our algorithm retains the main virtues of the Expectation-Maximization (EM) algorithm—its guarantee of monotonic im- provemen t, and its absence of tuning parameters—with the added advantage of optimizing a discriminative objective function. The algorithm reduces as a spe- cial case to the method of generalized iterative scaling for log-linear models. The learning rate of the algorithm is controlled by the sparseness of the training data. We use the method of nonnegative matrix factorization (NMF) to discover sparse dist ributed representations of the data. This form of feature selection greatly acce lerates learning and makes the algorithm practical on large problems. Ex- perime nts show that discriminatively trained mixture models lead to much better classi ■cation than comparably sized models trained by EM.

****

The Infinite Hidden Markov Model

Matthew Beal, Zoubin Ghahramani, Carl Rasmussen

We show that it is possible to extend hidden Markov models to have a countably i n■nite number of hidden states. By using the theory of Dirichlet processes we ca n implicitly integrate out the in■nitely many transition parameters, leaving onl y three hyperparameters which can be learned from data. These three hyperparamet ers de■ne a hierarchical Dirichlet process capable of capturing a rich set of tr ansition dynamics. The three hyperparameters control the time scale of the dynam ics, the sparsity of the underlying state-transition matrix, and the expected nu m- ber of distinct hidden states in a ■nite sequence. In this framework it is al so natural to allow the alphabet of emitted symbols to be in■nite— consider, for example, symbols being possible words appearing in En- glish text.

****

EM-DD: An Improved Multiple-Instance Learning Technique

Qi Zhang, Sally Goldman

We present a new multiple-instance (MI) learning technique (EM(cid:173) D D) that combines EM with the diverse density (DD) algorithm. EM-DD is a general-purpose MI algorithm that can be applied with boolean or real-value labels and makes real-value predictions. On the boolean Musk benchmarks, the EM-DD algorithm without any tuning significantly outperforms all previ ous algorithms. EM-DD is relatively insensitive to the number of relevant attributes in the data set and scales up well to large bag sizes. Fu rthermore, EM(cid:173) DD provides a new framework for MI learning, in which the MI problem is converted to a single-instance setting by us ing EM to estimate the instance responsible for the label of the bag.

****

Fast Parameter Estimation Using Green's Functions

K. Wong, F. Li

We propose a method for the fast estimation of hyperparameters in large networks, based on the linear response relation in the cav(cid:173) ity method , and an empirical measurement of the Green's func(cid:173) tion. Simul ation results show that it is efficient and precise, when compared with cr oss-validation and other techniques which require matrix inversion.

****

Information-Geometric Decomposition in Spike Analysis

Hiroyuki Nakahara, Shun-ichi Amari

We present an information-geometric measure to systematically investigate neuronal firing patterns, taking account not only of the second-order but also of higher-order interactions. We begin with the case of two ne urons for illustration and show how to test whether or not any pairwise corre lation in one period is significantly different from that in the other period. In order to test such a hy(cid:173) pothesis of different firing rates, t he correlation term needs to be singled out 'orthogonally' to the firing rat es, where the null hypoth(cid:173) esis might not be of independent firing. Thi s method is also shown to directly associate neural firing with behavior via their mutual information, which is decomposed into two types of information, conveyed by mean firing rate and coincident firing, respec

tively. Then, we show that these results, using the 'orthogonal' decompo(cid:173) sition, are naturally extended to the case of three neurons and n neurons in general.

****

Learning Body Pose via Specialized Maps

Rómer Rosales, Stan Sclaroff

A nonlinear supervised learning model, the Specialized Mappings Architectu re (SMA), is described and applied to the estimation of human body pose f rom monocular images. The SMA consists of several specialized forward map ping functions and an inverse map(cid:173) ping function. Each specialized f unction maps certain domains of the input space (image features) onto the output space (body pose parameters). The key algorithmic problems faced are those of learning the specialized domains and mapping functions in an o p(cid:173) timal way, as well as performing inference given inputs and knowl(c id:173) edge of the inverse function. Solutions to these problems employ the EM algorithm and alternating choices of conditional indepen(cid:173) de nce assumptions. Performance of the approach is evaluated with synthetic and real video sequences of human motion.

****

Citcuits for VLSI Implementation of Temporally Asymmetric Hebbian Learning

A. Bofill, D. Thompson, Alan Murray

Experimental data has shown that synaptic strength modification in some typ es of biological neurons depends upon precise spike tim(cid:173) ing difference s between presynaptic and postsynaptic spikes. Sev(cid:173) eral temporal ly-asymmetric Hebbian learning rules motivated by this data have been p roposed. We argue that such learning rules are suitable to analog VLS I implementation. We describe an eas(cid:173) ily tunable circuit to modi fy the weight of a silicon spiking neuron according to those learning rules. Test results from the fabrication of the circuit using a O.6J.lm CMOS proc ess are given.

****

Model Based Population Tracking and Automatic Detection of Distribution Changes

Igor Cadez, P. S. Bradley

Probabilistic mixture models are used for a broad range of data anal- ysis tasks such as clustering, classi■cation, predictive modeling, etc. Due to their inher ent probabilistic nature, mixture models can easily be combined with other proba bilistic or non-probabilistic techniques thus forming more complex data analysis systems. In the case of online data (where there is a stream of data available) models can be constantly up- dated to re■ect the most current distribution of t he incoming data. How- ever, in many business applications the models themselves represent a parsimonious summary of the data and therefore it is not desirable to change models frequently, much less with every new data point. In such a fram ework it becomes crucial to track the applicability of the mixture model and det ect the point in time when the model fails to adequately represent the data. In this paper we formulate the problem of change detection and propose a principled solution. Empirical results over both synthetic and real-life data sets are pre sented.

****

Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering

Mikhail Belkin, Partha Niyogi

Drawing on the correspondence between the graph Laplacian, the Laplace-B eltrami operator on a manifold , and the connections to the heat equatio n , we propose a geometrically motivated algorithm for constructing a repres entation for data sampled from a low di(cid:173) mensional manifold embedde d in a higher dimensional space. The algorithm provides a computational ly efficient approach to non(cid:173) linear dimensionality reduction tha t has locality preserving prop(cid:173) erties and a natural connection to clustering. Several applications are considered.

****

A Bayesian Model Predicts Human Parse Preference and Reading Times in Sentence P

rocessing
S. Narayanan, Daniel Jurafsky
Narayanan and Jurafsky (1998) proposed that human language compre- hension can b e modeled by treating human comprehenders as Bayesian reasoners, and modeling th e comprehension process with Bayesian de- cision trees. In this paper we extend the Narayanan and Jurafsky model to make further predictions about reading time given the probability of difference parses or interpretations, and test the mode l against reading time data from a psycholinguistic experiment.
****

MIME: Mutual Information Minimization and Entropy Maximization for Bayesian Beli ef Propagation
Anand Rangarajan, Alan L. Yuille
Bayesian belief propagation in graphical models has been recently shown to have very close ties to inference methods based in statis- tical physics. After Yedid ia et al. demonstrated that belief prop- agation (cid:12)xed points correspond t o extrema of the so-called Bethe free energy, Yuille derived a double loop algor ithm that is guar- anteed to converge to a local minimum of the Bethe free energ y. Yuille's algorithm is based on a certain decomposition of the Bethe free ener gy and he mentions that other decompositions are possi- ble and may even be frui tful. In the present work, we begin with the Bethe free energy and show that it has a principled interpre- tation as pairwise mutual information minimization an d marginal entropy maximization (MIME). Next, we construct a family of free ener gy functions from a spectrum of decompositions of the original Bethe free energy . For each free energy in this family, we develop a new algorithm that is guaran teed to converge to a local min- imum. Preliminary computer simulations are in a greement with this theoretical development.
****

Rates of Convergence of Performance Gradient Estimates Using Function Approximat ion and Bias in Reinforcement Learning
Gregory Grudic, Lyle Ungar

****

A Generalization of Principal Components Analysis to the Exponential Family
Michael Collins, S. Dasgupta, Robert E. Schapire
Principal component analysis (PCA) is a commonly applied technique for dimension ality reduction. PCA implicitly minimizes a squared loss function, which may be inappropriate for data that is not real-valued, such as binary-valued data. This paper draws on ideas from the Exponen- tial family, Generalized linear models, and Bregman distances, to give a generalization of PCA to loss functions that we argue are better suited to other data types. We describe algorithms for minimiz ing the loss func- tions, and give examples on simulated data.
****

Minimax Probability Machine
Gert Lanckriet, Laurent Ghaoui, Chiranjib Bhattacharyya, Michael Jordan
When constructing a classifier, the probability of correct classifi(cid:173) cation of future data points should be maximized. In the current pape r this desideratum is translated in a very direct way into an optimiz ation problem, which is solved using methods from con(cid:173) vex opti mization. We also show how to exploit Mercer kernels in this setting to obtain nonlinear decision boundaries. A worst-case bound on the probabili ty of misclassification of future data is ob(cid:173) tained explicitly.
****

Duality, Geometry, and Support Vector Regression
J. Bi, Kristin Bennett
We develop an intuitive geometric framework for support vector regression (SVR). By examining when (cid:15)-tubes exist, we show that SVR can be regarded as a c lassi(cid:12)cation problem in the dual space. Hard and soft (cid:15)-tubes are constructed by separating the convex or reduced convex hulls respectively of the training data with the response variable shifted up and down by (cid:15). A nov el SVR model is proposed based on choosing the max-margin plane between the two

shifted datasets. Maximizing the margin corresponds to shrinking the e(cid:11)ective (cid:15)-tube. In the proposed approach the e(cid:11)ects of the choices of all parameters become clear geometrically.

****

Blind Source Separation via Multinode Sparse Representation

Michael Zibulevsky, Pavel Kisilev, Yehoshua Zeevi, Barak Pearlmutter

We consider a problem of blind source separation from a set of instan(cid:173) taneous linear mixtures, where the mixing matrix is unknown. It was discovered recently, that exploiting the sparsity of sources in an appro(cid:173) priate representation according to some signal dictionary, dramatically improves the quality of separation. In this work we use the property of multi scale transforms, such as wavelet or wavelet packets, to decompose signals into sets of local features with various degrees of sparsity. We use this intrinsic property for selecting the best (most sparse) subsets of features for further separation. The performance of the algorithm is ver(cid:173) ified on noise-free and noisy data. Experiments with simulated signals, musical sounds and images demonstrate significant improvement of sep(cid:173) aration quality over previously reported results.

****

Estimating the Reliability of ICA Projections

Frank Meinecke, Andreas Ziehe, Motoaki Kawanabe, Klaus-Robert Müller

When applying unsupervised learning techniques like ICA or tem(cid:173) poral decorrelation, a key question is whether the discovered pro(cid:173) jections are reliable. In other words: can we give error bars or can we assess the quality of our separation? We use resampling meth(cid:173) ods to tackle these questions and show experimentally that our proposed variance estimations are strongly correlated to the sepa(cid:173) ration error. We demonstrate that this reliability estimation can be used to choose the appropriate ICA-model, to enhance signifi(cid:173) cantly the separation performance, and, most important, to mark the components that have a actual physical meaning. Application to 49-channel-data from an magneto encephalography (MEG) ex(cid:173) periment underlines the usefulness of our approach.

****

Learning from Infinite Data in Finite Time

Pedro Domingos, Geoff Hulten

We propose the following general method for scaling learning algorithms to arbitrarily large data sets. Consider the model Mii learned by the algorithm using ni examples in step i (ii = (nl , ... ,nm)) , and the model Moo that would be learned using in(cid:173) finite examples. Upper-bound the loss L(Mii' M oo) between them as a function of ii, and then minimize the algorithm's time com(cid:173) plexity f(ii) subject to the constraint that L(Moo , Mii) be at most f with probability at most 8. We apply this method to the EM algorithm for mixtures of Gaussians. Preliminary experiments on a series of large data sets provide evidence of the potential of this approach.

****

The g Factor: Relating Distributions on Features to Distributions on Images

James Coughlan, Alan L. Yuille

We describe the g-factor, which relates probability distributions on image features to distributions on the images themselves. The g-factor depends only on our choice of features and lattice quanti(cid:173) zation and is independent of the training image data. We illustrate the importance of the g-factor by analyzing how the parameters of Markov Random Field (i.e. Gibbs or log-linear) probability models of images are learned from data by maximum likelihood estimation. In particular, we study homogeneous MRF models which learn im(cid:173) age distributions in terms of clique potentials corresponding to fea(cid:173) ture histogram statistics (d. Minimax Entropy Learning (MEL) by Zhu, Wu and Mumford 1997 [11]) . We first use our analysis of the g-factor to determine when the clique potentials decouple for different features . Second, we show that clique potentials can be computed

analytically by approximating the g-factor. Third, we demonstrate a connection between this approximation and the Generalized Iterative Scaling algorithm (GIS), due to Darroch and Ratcliff 1972 [2], for calculating potentials. This connection en(cid:173) ables us to use GIS to improve our multinomial approximation, using Bethe-Kikuchi[8] approximations to simplify the GIS proce(cid:173) dure. We support our analysis by computer simulations.

****

Distribution of Mutual Information

Marcus Hutter

The mutual information of two random variables z and J with joint probabilities {7rij} is commonly used in learning Bayesian nets as well as in many other fields. The chances 7rij are usually estimated by the empirical sampling frequency nij In leading to a point es(cid:173) timate J(nij In) for the mutual information. To answer questions like "is J (nij In) consistent with zero?" or "what is the probability that the true mutual information is much larger than the point es(cid:173) timate?" one has to go beyond the point estimate. In the Bayesian framework one can answer these questions by utilizing a (second order) prior distribution p(7r) comprising prior information about 7r. From the prior p(7r) one can compute the posterior p(7rln), from which the distribution p(Iln) of the mutual information can be cal(cid:173) culated. We derive reliable and quickly computable approximations for p(Iln). We concentrate on the mean, variance, skewness, and kurtosis, and non-informative priors. For the mean we also give an exact expression. Numerical issues and the range of validity are discussed.

****

A Maximum-Likelihood Approach to Modeling Multisensory Enhancement

H. Colonius, A. Diederich

Multisensory response enhancement (MRE) is the augmentation of the response of a neuron to sensory input of one modality by si(cid:173) multaneous input from another modality. The maximum likelihood (ML) model presented here modifies the Bayesian model for MRE (Anastasio et al.) by incorporating a decision strategy to maximize the number of correct decisions. Thus the ML model can also deal with the important tasks of stimulus discrimination and identifi(cid:173) cation in the presence of incongruent visual and auditory cues. It accounts for the inverse effectiveness observed in neurophysiolog(cid:173) ical recording data, and it predicts a functional relation between uni- and bimodal levels of discriminability that is testable both in neurophysiological and behavioral experiments.

****

Model-Free Least-Squares Policy Iteration

Michail G. Lagoudakis, Ronald Parr

We propose a new approach to reinforcement learning which combines least squares function approximation with policy iteration. Our method is model-free and completely off policy. We are motivated by the least squares temporal difference learning algorithm (LSTD), which is known for its ef■cient use of sample experiences compared to pure temporal difference algorithms. LSTD is ideal for prediction problems, however it heretofore has not had a straightforward application to control problems. Moreover, approximations learned by LSTD are strongly in■uenced by the visitation distribution over states. Our new algorithm, Least Squares Policy Iteration (LSPI) addresses these issues. The result is an off-policy method which can use (or reuse) data collected from any source. We have tested LSPI on several problems, including a bicycle simulator in which it learns to guide the bicycle to a goal ef■ciently by merely observing a relatively small number of completely random trials.

****