Decoding V1 Neuronal Activity using Particle Filtering with Volterra Kernels
Ryan Kelly, Tai Sing Lee
Decoding is a strategy that allows us to assess the amount of information neuron s can provide about certain aspects of the visual scene. In this study, we devel op a method based on Bayesian sequential updating and the particle ■ltering algo rithm to decode the activity of V1 neurons in awake monkeys. A distinction in ou r method is the use of Volterra ker- nels to ■lter the particles, which live in a high dimensional space. This parametric Bayesian decoding scheme is compared t o the optimal linear decoder and is shown to work consistently better than the l inear optimal decoder. Interestingly, our results suggest that for decoding in r eal time, spike trains of as few as 10 independent but similar neurons would be suf■cient for decoding a critical scene variable in a particular class of visual stimuli. The reconstructed variable can predict the neural activity about as we ll as the actual signal with respect to the Volterra kernels.
****

Linear Response for Approximate Inference
Max Welling, Yee Teh
Belief propagation on cyclic graphs is an ef■cient algorithm for comput- ing app roximate marginal probability distributions over single nodes and neighboring no des in the graph. In this paper we propose two new al- gorithms for approximatin g joint probabilities of arbitrary pairs of nodes and prove a number of desirabl e properties that these estimates ful■ll. The ■rst algorithm is a propagation al gorithm which is shown to con- verge if belief propagation converges to a stable ■xed point. The second algorithm is based on matrix inversion. Experiments comp are a number of competing methods.
****

Wormholes Improve Contrastive Divergence
Max Welling, Andriy Mnih, Geoffrey E. Hinton
In models that de■ne probabilities via energies, maximum likelihood learning typ ically involves using Markov Chain Monte Carlo to sample from the model’s distri bution. If the Markov chain is started at the data distribution, learning often works well even if the chain is only run for a few time steps [3]. But if the da ta distribution contains modes separated by regions of very low density, brief M CMC will not ensure that different modes have the correct relative energies beca use it cannot move particles from one mode to another. We show how to improve br ief MCMC by allowing long-range moves that are suggested by the data distributio n. If the model is approximately correct, these long-range moves have a reasonab le acceptance rate.
****

Probability Estimates for Multi-Class Classification by Pairwise Coupling
Ting-fan Wu, Chih-jen Lin, Ruby Weng
Pairwise coupling is a popular multi-class classi■cation method that combines to gether all pairwise comparisons for each pair of classes. This paper presents tw o approaches for obtaining class probabilities. Both methods can be reduced to l inear systems and are easy to implement. We show conceptually and experimentally that the proposed approaches are more stable than two existing popular methods: voting and [3].
****

Information Dynamics and Emergent Computation in Recurrent Circuits of Spiking N eurons
Thomas Natschläger, Wolfgang Maass
We employ an ef■cient method using Bayesian and linear classi■ers for analyzing the dynamics of information in high-dimensional states of generic cortical micro circuit models. It is shown that such recurrent cir- cuits of spiking neurons ha ve an inherent capability to carry out rapid computations on complex spike patte rns, merging information contained in the order of spike arrival with previously acquired context information.
****

Mutual Boosting for Contextual Inference
Michael Fink, Pietro Perona

Mutual Boosting is a method aimed at incorporating contextual informati on to augment object detection. When multiple detectors of objects and parts are trained in parallel using AdaBoost [1], object detectors mig ht use the remaining intermediate detectors to enrich the weak learner set. This method generalizes the efficient features suggested by Viola and Jones thus enabling information inference between parts and objects in a co mpositional hierarchy. In our experiments eye-, nose-, mouth- and face detectors are trained using the Mutual Boosting framework. Results show the method outperforms applications overlooking contextual information. We su ggest that achieving contextual integration is a step toward human-like dete ction capabilities.

****

On the Dynamics of Boosting

Cynthia Rudin, Ingrid Daubechies, Robert E. Schapire

In order to understand AdaBoost’s dynamics, especially its ability to maximize m argins, we derive an associated simpli■ed nonlinear iterated map and analyze its behavior in low-dimensional cases. We ■nd stable cycles for these cases, which can explicitly be used to solve for Ada- Boost’s output. By considering AdaBoost as a dynamical system, we are able to prove R¨atsch and Warmuth’s conjecture th at AdaBoost may fail to converge to a maximal-margin combined classi■er when giv en a ‘non- optimal’ weak learning algorithm. AdaBoost is known to be a coordinat e descent method, but other known algorithms that explicitly aim to max- imize t he margin (such as AdaBoost⁄ and arc-gv) are not. We consider a differentiable f unction for which coordinate ascent will yield a maxi- mum margin solution. We t hen make a simple approximation to derive a new boosting algorithm whose updates are slightly more aggressive than those of arc-gv.

****

Distributed Optimization in Adaptive Networks

Ciamac C. Moallemi, Benjamin Roy

We develop a protocol for optimizing dynamic behavior of a network of simple ele ctronic components, such as a sensor network, an ad hoc network of mobile device s, or a network of communication switches. This protocol requires only local com munication and simple computa- tions which are distributed among devices. The pr otocol is scalable to large networks. As a motivating example, we discuss a prob lem involv- ing optimization of power consumption, delay, and buffer over■ow in a sensor network. Our approach builds on policy gradient methods for optimizatio n of Markov decision processes. The protocol can be viewed as an extension of po licy gradient methods to a context involving a team of agents op- timizing aggre gate performance through asynchronous distributed com- munication and computatio n. We establish that the dynamics of the pro- tocol approximate the solution to an ordinary differential equation that follows the gradient of the performance o bjective.

****

Statistical Debugging of Sampled Programs

Alice Zheng, Michael Jordan, Ben Liblit, Alex Aiken

We present a novel strategy for automatically debugging programs given sampled d ata from thousands of actual user runs. Our goal is to pinpoint those features t hat are most correlated with crashes. This is accomplished by maximizing an appr opriately de■ned utility function. It has analogies with intuitive debugging heu ristics, and, as we demonstrate, is able to deal with various types of bugs that occur in real programs.

****

A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications

Pedro Moreno, Purdy Ho, Nuno Vasconcelos

Over the last years signi■cant efforts have been made to develop kernels that ca n be applied to sequence data such as DNA, text, speech, video and images. The F isher Kernel and similar variants have been suggested as good ways to combine an underlying generative model in the feature space and discriminant classi■ers su ch as SVM’s. In this paper we sug- gest an alternative procedure to the Fisher k

ernel for systematically ■nd- ing kernel functions that naturally handle variabl e length sequence data in multimedia domains. In particular for domains such as speech and images we explore the use of kernel functions that take full advantag e of well known probabilistic models such as Gaussian Mixtures and sin- gle full covariance Gaussian models. We derive a kernel distance based on the Kullback-L eibler (KL) divergence between generative models. In effect our approach combine s the best of both generative and discrim- inative methods and replaces the stan dard SVM kernels. We perform experiments on speaker identi■cation/veri■cation an d image classi■ca- tion tasks and show that these new kernels have the best perf ormance in speaker veri■cation and mostly outperform the Fisher kernel based SVM ’s and the generative classi■ers in speaker identi■cation and image classi■catio n.

****

A Model for Learning the Semantics of Pictures

Victor Lavrenko, R. Manmatha, Jiwoon Jeon

We propose an approach to learning the semantics of images which al- lows us to automatically annotate an image with keywords and to retrieve images based on te xt queries. We do this using a formalism that models the generation of annotated images. We assume that every image is di- vided into regions, each described by a continuous-valued feature vector. Given a training set of images with annotat ions, we compute a joint prob- abilistic model of image features and words which allow us to predict the probability of generating a word given the image region s. This may be used to automatically annotate and retrieve images given a word a s a query. Experiments show that our model signi■cantly outperforms the best of the previously reported results on the tasks of automatic image annotation and r etrieval.

****

The Diffusion-Limited Biochemical Signal-Relay Channel

Peter Thomas, Donald Spencer, Sierra Hampton, Peter Park, Joseph Zurkus are

****

Margin Maximizing Loss Functions

Saharon Rosset, Ji Zhu, Trevor Hastie

Margin maximizing properties play an important role in the analysis of classi£- cation models, such as boosting and support vector machines. Margin maximiza- ti on is theoretically interesting because it facilitates generalization error anal ysis, and practically interesting because it presents a clear geometric interpre tation of the models being built. We formulate and prove a suf£cient condition f or the solutions of regularized loss functions to converge to margin maximizing separa- tors, as the regularization vanishes. This condition covers the hinge lo ss of SVM, the exponential loss of AdaBoost and logistic regression loss. We als o generalize it to multi-class classi£cation problems, and present margin maximi zing multi- class versions of logistic regression and support vector machines.

****

Inferring State Sequences for Non-linear Systems with Embedded Hidden Markov Mod els

Radford Neal, Matthew Beal, Sam Roweis

We describe a Markov chain method for sampling from the distribution of the hidd en state sequence in a non-linear dynamical system, given a sequence of observat ions. This method updates all states in the sequence simultaneously using an emb edded Hidden Markov Model (HMM). An update begins with the creation of “pools” o f candidate states at each time. We then de■ne an embedded HMM whose states are indexes within these pools. Using a forward-backward dynamic programming algo- r ithm, we can ef■ciently choose a state sequence with the appropriate probabiliti es from the exponentially large number of state sequences that pass through stat es in these pools. We illustrate the method in a simple one-dimensional example, and in an example showing how an embed- ded HMM can be used to in effect discre tize the state space without any discretization error. We also compare the embed ded HMM to a particle smoother on a more substantial problem of inferring human motion from 2D traces of markers.

****
Self-calibrating Probability Forecasting
Vladimir Vovk, Glenn Shafer, Ilia Nouretdinov
In the problem of probability forecasting the learner’s goal is to output, given a training set and a new object, a suitable probability measure on the possible values of the new object’s label. An on-line algorithm for probability forecasting is said to be well-calibrated if the probabilities it outputs agree with the observed frequencies. We give a natural non- asymptotic formalization of the notion of well-calibratedness, which we then study under the assumption of randomness (the object/label pairs are independent and identically distributed). It turns out that, although no probability forecasting algorithm is automatically well-calibrated in our sense, there exists a wide class of algorithms for “multiprobability forecasting” (such algorithms are allowed to output a set, ideally very narrow, of probability measures) which satisfy this property; we call the algorithms in this class “Venn probability machines”. Our experimental results demonstrate that a 1-Nearest Neighbor Venn probability machine performs reasonably well on a standard benchmark data set, and one of our theoretical results asserts that a simple Venn probability machine asymptotically approaches the true conditional probabilities regardless, and without knowledge, of the true probability measure generating the examples.
****
Synchrony Detection by Analogue VLSI Neurons with Bimodal STDP Synapses
Adria Bofill-i-petit, Alan Murray
We present test results from spike-timing correlation learning ex- periments carried out with silicon neurons with STDP (Spike Tim- ing Dependent Plasticity) synapses. The weight change scheme of the STDP synapses can be set to either weight-independent or weight-dependent mode. We present results that characterise the learning window implemented for both modes of operation. When presented with spike trains with di(cid:11)erent types of synchronisation the neurons develop bimodal weight distributions. We also show that a 2-layered network of silicon spiking neurons with STDP synapses can perform hierarchical synchrony detection.
****
Semi-supervised Protein Classification Using Cluster Kernels
Jason Weston, Dengyong Zhou, André Elisseeff, William Noble, Christina Leslie
A key issue in supervised protein classi■cation is the representation of in- put sequences of amino acids. Recent work using string kernels for pro- tein data has achieved state-of-the-art classi■cation performance. How- ever, such representations are based only on labeled data — examples with known 3D structures, organized into structural classes — while in practice, unlabeled data is far more plentiful. In this work, we de- velop simple and scalable cluster kernel techniques for incorporating un- labeled data into the representation of protein sequences. We show that our methods greatly improve the classi■cation performance of string ker- nels and outperform standard approaches for using unlabeled data, such as adding close homologs of the positive examples to the training data. We achieve equal or superior performance to previously presented cluster kernel methods while achieving far greater computational ef■ciency.
****
When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?
David Donoho, Victoria Stodden
We interpret non-negative matrix factorization geometrically, as the problem of ■nding a simplicial cone which contains a cloud of data points and which is contained in the positive orthant. We show that under certain conditions, basically requiring that some of the data are spread across the faces of the positive orthant, there is a unique such simpli- cial cone. We give examples of synthetic image articulation databases which obey these conditions; these require separated support and facto- rial sampling. For such databases there is a generative model in terms of ‘parts’ and NMF correctly identi■es the ‘parts’. We show that our theoretical results are predictive of the performance of published NMF code, by running the published algorithms on one of our synthetic image articulation databa

ses.
****
A Biologically Plausible Algorithm for Reinforcement-shaped Representational Lea rning
Maneesh Sahani
Signi■cant plasticity in sensory cortical representations can be driven in matur e animals either by behavioural tasks that pair sensory stimuli with reinforceme nt, or by electrophysiological experiments that pair sensory input with direct s timulation of neuromodulatory nuclei, but usually not by sensory stimuli present ed alone. Biologically motivated theories of representational learning, however, have tended to focus on unsupervised mechanisms, which may play a signi■cant ro le on evolutionary or devel- opmental timescales, but which neglect this essenti al role of reinforce- ment in adult plasticity. By contrast, theoretical reinfor cement learning has generally dealt with the acquisition of optimal policies for action in an uncertain world, rather than with the concurrent shaping of sensor y representations. This paper develops a framework for representational learning which builds on the relative success of unsupervised generative- modelling acco unts of cortical encodings to incorporate the effects of reinforcement in a biol ogically plausible way.
****
Online Classification on a Budget
Koby Crammer, Jaz Kandola, Yoram Singer
Online algorithms for classi■cation often require vast amounts of mem- ory and c omputation time when employed in conjunction with kernel functions. In this pape r we describe and analyze a simple approach for an on-the-■y reduction of the nu mber of past examples used for prediction. Experiments performed with real datas ets show that using the proposed algorithmic approach with a single epoch is com petitive with the sup- port vector machine (SVM) although the latter, being a ba tch algorithm, accesses each training example multiple times.
****
Pairwise Clustering and Graphical Models
Noam Shental, Assaf Zomet, Tomer Hertz, Yair Weiss
Signi(cid:2)cant progress in clustering has been achieved by algorithms that are based on pairwise af(cid:2)nities between the datapoints. In particular, spectr al clustering methods have the advantage of being able to divide arbitrarily sha ped clusters and are based on ef(cid:2)cient eigenvector calcu- lations. Howeve r, spectral methods lack a straightforward probabilistic interpretation which mak es it dif(cid:2)cult to automatically set parameters us- ing training data. In t his paper we use the previously proposed typical cut framework for pairwise clus tering. We show an equivalence between calculating the typical cut and inference in an undirected graphical model. We show that for clustering problems with hun dreds of datapoints exact inference may still be possible. For more complicated datasets, we show that loopy be- lief propagation (BP) and generalized belief pr opagation (GBP) can give excellent results on challenging clustering problems. W e also use graph- ical models to derive a learning algorithm for af(cid:2)nity m atrices based on labeled data.
****
Link Prediction in Relational Data
Ben Taskar, Ming-fai Wong, Pieter Abbeel, Daphne Koller
Many real-world domains are relational in nature, consisting of a set of objects related to each other in complex ways. This paper focuses on predicting the exi stence and the type of links between entities in such domains. We apply the rela tional Markov network framework of Taskar et al. to de■ne a joint probabilis- ti c model over the entire link graph — entity attributes and links. The applicatio n of the RMN algorithm to this task requires the de■nition of probabilistic patt erns over subgraph structures. We apply this method to two new relational datase ts, one involving university webpages, and the other a social network. We show t hat the collective classi■cation approach of RMNs, and the introduction of subgr aph patterns over link labels, provide signi■cant improvements in accuracy over ■at classi■cation, which attempts to predict each link in isolation.

****
Human and Ideal Observers for Detecting Image Curves
Fang Fang, Daniel Kersten, Paul R. Schrater, Alan L. Yuille
This paper compares the ability of human observers to detect target im- age curv es with that of an ideal observer. The target curves are sam- pled from a genera tive model which speci■es (probabilistically) the ge- ometry and local intensity properties of the curve. The ideal observer performs Bayesian inference on the generative model using MAP esti- mation. Varying the probability model for the c urve geometry enables us investigate whether human performance is best for targe t curves that obey speci■c shape statistics, in particular those observed on nat ural shapes. Experiments are performed with data on both rectangular and hexagon al lattices. Our results show that human observers' performance approaches that of the ideal observer and are, in general, closest to the ideal for con- ditions where the target curve tends to be straight or similar to natural statistics on curves. This suggests a bias of human observers towards straight curves and nat ural statistics.
****
Linear Program Approximations for Factored Continuous-State Markov Decision Proc esses
Milos Hauskrecht, Branislav Kveton
Approximate linear programming (ALP) has emerged recently as one of the most pro mising methods for solving complex factored MDPs with (cid:2)nite state spaces. In this work we show that ALP solutions are not limited only to MDPs with (cid:2)nite state spaces, but that they can also be applied successfully to factored c ontinuous-state MDPs (CMDPs). We show how one can build an ALP-based approximati on for such a model and contrast it to existing solution methods. We argue that this approach offers a robust alternative for solving high dimensional continuou s-state space problems. The point is supported by experiments on three CMDP prob lems with 24-25 continuous state factors.
****
Perception of the Structure of the Physical World Using Unknown Multimodal Senso rs and Effectors
D. Philipona, J.k. O'regan, J.-p. Nadal, Olivier Coenen
Is there a way for an algorithm linked to an unknown body to infer by itself inf ormation about this body and the world it is in? Taking the case of space for ex ample, is there a way for this algorithm to realize that its body is in a three dimensional world? Is it possible for this algorithm to discover how to move in a straight line? And more basically: do these questions make any sense at all gi ven that the algorithm only has access to the very high-dimensional data consist ing of its sensory inputs and motor outputs? We demonstrate in this article how these questions can be given a positive answer. We show that it is possible to m ake an algorithm that, by ana- lyzing the law that links its motor outputs to it s sensory inputs, discovers information about the structure of the world regardl ess of the devices constituting the body it is linked to. We present results fro m simulations demonstrating a way to issue motor orders resulting in "fundamenta l" movements of the body as regards the structure of the physical world.
****
Log-Linear Models for Label Ranking
Ofer Dekel, Yoram Singer, Christopher D. Manning
Label ranking is the task of inferring a total order over a prede■ned set of lab els for each given instance. We present a general framework for batch learning o f label ranking functions from supervised data. We assume that each instance in the training data is associated with a list of preferences over the label-set, h owever we do not assume that this list is either com- plete or consistent. This enables us to accommodate a variety of ranking problems. In contrast to the gene ral form of the supervision, our goal is to learn a ranking function that induce s a total order over the entire set of labels. Special cases of our setting are multilabel categorization and hierarchical classi■cation. We present a general b oosting-based learning algorithm for the label ranking problem and prove a lower bound on the progress of each boosting iteration. The applicability of our appr

oach is demonstrated with a set of experiments on a large-scale text corpus.
****

Learning the k in k-means
Greg Hamerly, Charles Elkan
When clustering a dataset, the right number k of clusters to use is often not ob vious, and choosing k automatically is a hard algorithmic prob- lem. In this pap er we present an improved algorithm for learning k while clustering. The G-means algorithm is based on a statistical test for the hypothesis that a subset of da ta follows a Gaussian distribution. G-means runs k-means with increasing k in a hierarchical fashion until the test ac- cepts the hypothesis that the data assig ned to each k-means center are Gaussian. Two key advantages are that the hypothe sis test does not limit the covariance of the data and does not compute a full c ovariance matrix. Additionally, G-means only requires one intuitive parameter, t he stand- ard statistical signi■cance level α. We present results from experimen ts showing that the algorithm works well, and better than a recent method based on the BIC penalty for model complexity. In these experiments, we show that the BIC is ineffective as a scoring function, since it does not penalize strongly en ough the model’s complexity.
****

Learning a Rare Event Detection Cascade by Direct Feature Selection
Jianxin Wu, James M. Rehg, Matthew Mullin
Face detection is a canonical example of a rare event detection prob- lem, in wh ich target patterns occur with much lower frequency than non- targets. Out of mi llions of face-sized windows in an input image, for ex- ample, only a few will t ypically contain a face. Viola and Jones recently proposed a cascade architectur e for face detection which successfully ad- dresses the rare event nature of the task. A central part of their method is a feature selection algorithm based on AdaBoost. We present a novel cascade learning algorithm based on forward feature selection which is two orders of magnitude faster than the Viola-Jones approach and yields classi■ers of equivalent quality. This faster method could be used f or more demanding classi■cation tasks, such as on-line learning.
****

Non-linear CCA and PCA by Alignment of Local Models
Jakob Verbeek, Sam Roweis, Nikos Vlassis
We propose a non-linear Canonical Correlation Analysis (CCA) method which works by coordinating or aligning mixtures of linear models. In the same way that CCA extends the idea of PCA, our work extends re- cent methods for non-linear dimens ionality reduction to the case where multiple embeddings of the same underlying low dimensional coordi- nates are observed, each lying on a different high dimen sional manifold. We also show that a special case of our method, when applied to only a single manifold, reduces to the Laplacian Eigenmaps algorithm. As with p revious alignment schemes, once the mixture models have been estimated, all of t he parameters of our model can be estimated in closed form without local optima in the learning. Experimental results illustrate the viability of the approach a s a non-linear extension of CCA.
****

Unsupervised Context Sensitive Language Acquisition from a Large Corpus
Zach Solan, David Horn, Eytan Ruppin, Shimon Edelman
We describe a pattern acquisition algorithm that learns, in an unsuper- vised fa shion, a streamlined representation of linguistic structures from a plain natura l-language corpus. This paper addresses the issues of learn- ing structured know ledge from a large-scale natural language data set, and of generalization to uns een text. The implemented algorithm repre- sents sentences as paths on a graph w hose vertices are words (or parts of words). Signi■cant patterns, determined by recursive context-sensitive statistical inference, form new vertices. Linguistic constructions are rep- resented by trees composed of signi■cant patterns and th eir associated equivalence classes. An input module allows the algorithm to be s ub- jected to a standard test of English as a Second Language (ESL) pro■- ciency . The results are encouraging: the model attains a level of per- formance consid ered to be “intermediate” for 9th-grade students, de- spite having been trained

on a corpus (CHILDES) containing transcribed speech of parents directed to small children.

****

Modeling User Rating Profiles For Collaborative Filtering

Benjamin M. Marlin

In this paper we present a generative latent variable model for rating-based collaborative (cid:12)ltering called the User Rating Pro(cid:12)le model (URP). The generative process which underlies URP is de- signed to produce complete user rating pro(cid:12)les, an assignment of one rating to each item for each user. Our model represents each user as a mixture of user attitudes, and the mixing proportions are distributed according to a Dirichlet random variable. The rating for each item is generated by selecting a user attitude for the item, and then selecting a rating according to the preference pattern associ- ated with that attitude. URP is related to several models including a multinomial mixture model, the aspect model [7], and LDA [1], but has clear advantages over each.

****

Learning a World Model and Planning with a Self-Organizing, Dynamic Neural System

Marc Toussaint

We present a connectionist architecture that can learn a model of the relations between perceptions and actions and use this model for be- havior planning. State representations are learned with a growing self- organizing layer which is directly coupled to a perception and a motor layer. Knowledge about possible state transitions is encoded in the lat- eral connectivity. Motor signals modulate this lateral connectivity and a dynamic ■eld on the layer organizes a planning process. All mecha- nisms are local and adaptation is based on Hebbian ideas. The model is continuous in the action, perception, and time domain.

****

Bias-Corrected Bootstrap and Model Uncertainty

Harald Steck, Tommi Jaakkola

The bootstrap has become a popular method for exploring model (structure) uncertainty. Our experiments with arti■cial and real- world data demonstrate that the graphs learned from bootstrap samples can be severely biased towards too complex graphical mod- els. Accounting for this bias is hence essential, e.g., when explor- ing model uncertainty. We ■nd that this bias is intimately tied to (well-known) spurious dependences induced by the bootstrap. The leading-order bias-correction equals one half of Akaike’s penalty for model complexity. We demonstrate the e■ect of this simple bias-correction in our experiments. We also relate this bias to the bias of the plug-in estimator for entropy, as well as to the di■er- ence between the expected test and training errors of a graphical model, which asymptotically equals Akaike’s penalty (rather than one half).

****

Nonlinear Processing in LGN Neurons

Vincent Bonin, Valerio Mante, Matteo Carandini

According to a widely held view, neurons in lateral geniculate nucleus (LGN) operate on visual stimuli in a linear fashion. There is ample evidence, however, that LGN responses are not entirely linear. To account for nonlinearities we propose a model that synthesizes more than 30 years of research in the field. Model neurons have a linear receptive field, and a nonlinear, divisive suppressive field. The suppressive field computes local root-mean- square contrast. To test this model we recorded responses from LGN of anesthetized paralyzed cats. We estimate model parameters from a basic set of measurements and show that the model can accurately predict responses to novel stimuli. The model might serve as the new standard model of LGN responses. It specifies how visual processing in LGN involves both linear filtering and divisive gain control.

****

Ranking on Data Manifolds

Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, Bernhard Schölkopf

The Google search engine has enjoyed huge success with its web page ranking algo rithm, which exploits global, rather than local, hyperlink structure of the web using random walks. Here we propose a simple universal ranking algorithm for dat a lying in the Euclidean space, such as text or image data. The core idea of our method is to rank the data with respect to the intrinsic manifold structure col lectively revealed by a great amount of data. Encouraging experimental results f rom synthetic, image, and text data illustrate the validity of our method.
****

Approximate Analytical Bootstrap Averages for Support Vector Classifiers
Dörthe Malzahn, Manfred Opper
We compute approximate analytical bootstrap averages for support vec- tor classi ■cation using a combination of the replica method of statistical physics and the TAP approach for approximate inference. We test our method on a few datasets an d compare it with exact averages obtained by extensive Monte-Carlo sampling.
****

An Infinity-sample Theory for Multi-category Large Margin Classification
Tong Zhang
The purpose of this paper is to investigate in■nity-sample properties of risk mi nimization based multi-category classi■cation methods. These methods can be cons idered as natural extensions to binary large margin classi■cation. We establish conditions that guarantee the in■nity-sample consistency of classi■ers obtained in the risk minimization framework. Examples are provided for two speci■c forms of the general formulation, which extend a number of known methods. Using these examples, we show that some risk minimization formulations can also be used to o b- tain conditional probability estimates for the underlying problem. Such condi tional probability information will be useful for statistical inferenc- ing task s beyond classi■cation.
****

A Neuromorphic Multi-chip Model of a Disparity Selective Complex Cell
Bertram Shi, Eric Tsang
The relative depth of objects causes small shifts in the left and right ret- ina l positions of these objects, called binocular disparity. Here, we describe a ne uromorphic implementation of a disparity selective com- plex cell using the bino cular energy model, which has been proposed to model the response of disparity s elective cells in the visual cortex. Our system consists of two silicon chips co ntaining spiking neurons with monocular Gabor-type spatial receptive fields (RF) and circuits that combine the spike outputs to compute a disparity selective co mplex cell response. The disparity selectivity of the cell can be adjusted by bo th position and phase shifts between the monocular RF profiles, which are both u sed in biology. Our neuromorphic system performs better with phase encoding, bec ause the relative responses of neurons tuned to dif- ferent disparities by phase shifts are better matched than the responses of neurons tuned by position shift s.
****

Robustness in Markov Decision Problems with Uncertain Transition Matrices
Arnab Nilim, Laurent Ghaoui
Optimal solutions to Markov Decision Problems (MDPs) are very sen- sitive with r espect to the state transition probabilities. In many practi- cal problems, the estimation of those probabilities is far from accurate. Hence, estimation errors are limiting factors in applying MDPs to real- world problems. We propose an al gorithm for solving ■nite-state and ■nite-action MDPs, where the solution is gua ranteed to be robust with respect to estimation errors on the state transition p robabilities. Our al- gorithm involves a statistically accurate yet numerically ef■cient repre- sentation of uncertainty, via Kullback-Leibler divergence bounds . The worst-case complexity of the robust algorithm is the same as the origi- na l Bellman recursion. Hence, robustness can be added at practically no extra comp uting cost.
****

Parameterized Novelty Detectors for Environmental Sensor Monitoring
Cynthia Archer, Todd Leen, António Baptista

As part of an environmental observation and forecasting system, sensors deployed in the Columbia RIver Estuary (CORIE) gather information on physical dynamics and changes in estuary habi- tat. Of these, salinity sensors are particularly susceptible to bio- fouling, which gradually degrades sensor response and corrupts crit- ical data. Automatic fault detectors have the capability to identify bio-fouling early and minimize data loss. Complicating the devel- opment of discriminatory classi(cid:12)ers is the scarcity of bio-fouling onset examples and the variability of the bio-fouling signature. To solve these problems, we take a novelty detection approach that incorporates a parameterized bio-fouling model. These detectors identify the occurrence of bio-fouling, and its onset time as reliably as human experts. Real-time detectors installed during the sum- mer of 2001 produced no false alarms, yet detected all episodes of sensor degradation before the (cid:12)eld sta(cid:11) scheduled these sensors for cleaning. From this initial deployment through February 2003, our bio-fouling detectors have essentially doubled the amount of useful data coming from the CORIE sensors.

****

Sensory Modality Segregation

Virginia Sa

Why are sensory modalities segregated the way they are? In this paper we show that sensory modalities are well designed for self-supervised cross-modal learning. Using the Minimizing-Disagreement algorithm on an unsupervised speech categorization task with visual (moving lips) and auditory (sound signal) inputs, we show that very informative auditory dimensions actually harm performance when moved to the visual side of the network. It is better to throw them away than to consider them part of the “visual input”. We explain this ■nding in terms of the statistical structure in sensory inputs.

****

Nonstationary Covariance Functions for Gaussian Process Regression

Christopher Paciorek, Mark Schervish

We introduce a class of nonstationary covariance functions for Gaussian process (GP) regression. Nonstationary covariance functions allow the model to adapt to functions whose smoothness varies with the inputs. The class includes a nonstationary version of the Matérn stationary co- variance, in which the differentiability of the regression function is con- trolled by a parameter, freeing one from ■xing the differentiability in advance. In experiments, the nonstationary GP regression model per- forms well when the input space is two or three dimensions, outperform- ing a neural network model and Bayesian free-knot spline models, and competitive with a Bayesian neural network, but is outperformed in one dimension by a state-of-the-art Bayesian free-knot spline model. The model readily generalizes to non-Gaussian data. Use of computational methods for speeding GP ■tting may allow for implementation of the method on larger datasets.

****

Algorithms for Interdependent Security Games

Michael Kearns, Luis E. Ortiz

nspired by events ranging from 9/11 to the collapse of the accounting ■rm Arthur Ander-
sen, economists Kunreuther and Heal [5] recently introduced an interesting game-theoretic
model for problems of interdependent security (IDS), in which a large number of players
must make individual investment decisions related to security — whether physical, ■nan-
cial, medical, or some other type — but in which the ultimate safety of each participant
may depend in a complex way on the actions of the entire population. A simple example is
the choice of whether to install a ■re sprinkler system in an individual condominium in a
large building. While such a system might greatly reduce the chances of the owner’s prop-

erty being destroyed by a ■re originating within their own unit, it might do lit tle or nothing
to reduce the chances of damage caused by ■res originating in other units (since sprinklers
can usually only douse small ■res early). If “enough” other unit owners have not made the
investment in sprinklers, it may be not cost-effective for any individual to do so.
****

How to Combine Expert (and Novice) Advice when Actions Impact the Environment?
Daniela de Farias, Nimrod Megiddo
The so-called “experts algorithms” constitute a methodology for choos- ing actio ns repeatedly, when the rewards depend both on the choice of action and on the u nknown current state of the environment. An experts algorithm has access to a se t of strategies (“experts”), each of which may recommend which action to choose. The algorithm learns how to com- bine the recommendations of individual experts so that, in the long run, for any ■xed sequence of states of the environment, i t does as well as the best expert would have done relative to the same sequence. This method- ology may not be suitable for situations where the evolution of st ates of the environment depends on past chosen actions, as is usually the case, for example, in a repeated non-zero-sum game. A new experts algorithm is present ed and analyzed in the context of re- peated games. It is shown that asymptotica lly, under certain conditions, it performs as well as the best available expert. This algorithm is quite different from previously proposed experts algorithms. It represents a shift from the paradigms of regret minimization and myopic optim iza- tion to consideration of the long-term effect of a player’s actions on the opponent’s actions or the environment. The importance of this shift is demonstra ted by the fact that this algorithm is capable of inducing co- operation in the repeated Prisoner’s Dilemma game, whereas previous experts algorithms converge t o the suboptimal non-cooperative play.
****

Reasoning about Time and Knowledge in Neural Symbolic Learning Systems
Artur Garcez, Luis Lamb
We show that temporal logic and combinations of temporal logics and modal logics of knowledge can be effectively represented in ar(cid:173) tificial neu ral networks. We present a Translation Algorithm from temporal rules to neural networks, and show that the networks compute a fixed-point se mantics of the rules. We also apply the translation to the muddy children puzzle, which has been used as a testbed for distributed multi-agent systems. We provide a complete solution to the puzzle with the use of simple neural net works, capa(cid:173) ble of reasoning about time and of knowledge acquisitio n through inductive learning.
****

Policy Search by Dynamic Programming
J. Bagnell, Sham M. Kakade, Jeff Schneider, Andrew Ng
We consider the policy search approach to reinforcement learning. We show that i f a “baseline distribution” is given (indicating roughly how often we expect a g ood policy to visit each state), then we can derive a policy search algorithm th at terminates in a ■nite number of steps, and for which we can provide non-trivi al performance guarantees. We also demonstrate this algorithm on several grid-wo rld POMDPs, a planar biped walking robot, and a double-pole balancing problem.
****

A Holistic Approach to Compositional Semantics: a connectionist model and robot experiments
Yuuya Sugita, Jun Tani
We present a novel connectionist model for acquiring the semantics of a simple l anguage through the behavioral experiences of a real robot. We focus on the “com positionality” of semantics, a fundamental character- istic of human language, w hich is the ability to understand the meaning of a sentence as a combination of the meanings of words. We also pay much attention to the “embodiment” of a robot

, which means that the robot should acquire semantics which matches its body, or sensory-motor system. The essential claim is that an embodied compositional semantic representation can be self-organized from generalized correspondences between sentences and behavioral patterns. This claim is examined and con■rmed through simple experiments in which a robot generates corre- sponding behaviors from unlearned sentences by analogy with the corre- spondences between learned sentences and behaviors.

****

Semidefinite Relaxations for Approximate Inference on Graphs with Cycles

Michael Jordan, Martin J. Wainwright

We present a new method for calculating approximate marginals for probability distributions de■ned by graphs with cycles, based on a Gaus- sian entropy bound combined with a semide■nite outer bound on the marginal polytope. This combination leads to a log-determinant max- imization problem that can be solved by ef■cient interior point meth- ods [8]. As with the Bethe approximation and its generalizations [12], the optimizing arguments of this problem can be taken as approximations to the exact marginals. In contrast to Bethe/Kikuchi approaches, our vari- ational problem is strictly convex and so has a unique global optimum. An additional desirable feature is that the value of the optimal solution is guaranteed to provide an upper bound on the log partition function. In experimental trials, the performance of the log-determinant relaxation is comparable to or better than the sum-product algorithm, and by a sub- stantial margin for certain problem classes. Finally, the zero-temperature limit of our log-determinant relaxation recovers a class of well-known semide■nite relaxations for integer programming [e.g., 3].

****

Discriminating Deformable Shape Classes

Salvador Ruiz-correa, Linda Shapiro, Marina Meila, Gabriel Berson

We present and empirically test a novel approach for categorizing 3-D free form ob- ject shapes represented by range data . In contrast to traditional surface-signature based systems that use alignment to match speci■c objects, we adapted the newly introduced symbolic-signature representation to classify deformable shapes [10]. Our approach con- structs an abstract description of shape classes using an ensemble of classi■ers that learn object class parts and their corresponding geometrical relationships from a set of numeric and symbolic descriptors. We used our classi■cation engine in a series of large scale dis- crimination experiments on two well-de■ned classes that share many common distinctive features. The experimental results suggest that our method outperforms traditional numeric signature-based methodologies. 1

****

Limiting Form of the Sample Covariance Eigenspectrum in PCA and Kernel PCA

David Hoyle, Magnus Rattray

We derive the limiting form of the eigenvalue spectrum for sample co- variance matrices produced from non-isotropic data. For the analysis of standard PCA we study the case where the data has increased variance along a small number of symmetry-breaking directions. The spectrum depends on the strength of the symmetry-breaking signals and on a pa- rameter (cid:11) which is the ratio of sample size to data dimension. Results are derived in the limit of large data dimension while keeping (cid:11) ■xed. As (cid:11) increases there are transitions in which delta functions emerge from the upper end of the bulk spectrum, corresponding to the symmetry-breaking directions in the data, and we calculate the bias in the corresponding eigenvalues. For kernel PCA the covariance matrix in feature space may contain symmetry-breaking structure even when the data components are independently distributed with equal variance. We show examples of phase-transition behaviour analogous to the PCA results in this case.

****

Auction Mechanism Design for Multi-Robot Coordination

Curt Bererton, Geoffrey J. Gordon, Sebastian Thrun

The design of cooperative multi-robot systems is a highly active research area in robotics. Two lines of research in particular have generated inter- est: the s

olution of large, weakly coupled MDPs, and the design and im- plementation of ma rket architectures. We propose a new algorithm which joins together these two li nes of research. For a class of coupled MDPs, our algorithm automatically design s a market architecture which causes a decentralized multi-robot system to conve rge to a consistent policy. We can show that this policy is the same as the one which would be produced by a particular centralized planning algorithm. We demon strate the new algorithm on three simulation examples: multi-robot towing, multi -robot path planning with a limited fuel resource, and coordinating behaviors in a game of paint ball.

****

Geometric Analysis of Constrained Curves

Anuj Srivastava, Washington Mio, Xiuwen Liu, Eric Klassen

We present a geometric approach to statistical shape analysis of closed curves i n images. The basic idea is to specify a space of closed curves satisfying given constraints, and exploit the differential geometry of this space to solve optim ization and inference problems. We demonstrate this approach by: (i) de■ning and computing statistics of observed shapes, (ii) de■ning and learning a parametric probability model on shape space, and (iii) designing a binary hypothesis test on this space.

****

A Low-Power Analog VLSI Visual Collision Detector

Reid Harrison

We have designed and tested a single-chip analog VLSI sensor that detects immi nent collisions by measuring radially expansive optic flow. The desig n of the chip is based on a model proposed to explain leg-extension behavior in flies during landing approaches. A new elementary motion detecto r (EMD) circuit was developed to measure optic flow. This EMD circuit models t he bandpass nature of large monopolar cells (LMCs) immediately postsynapt ic to photoreceptors in the fly visual system. A 16 × 16 array of 2-D motion detectors was fabricated on a 2.24 mm × 2.24 mm die in a standard 0.5-μm CMOS process. The chip consumes 140 μW of power from a 5 V s upply. With the addition of wide-angle optics, the sensor is able to detect co llisions around 500 ms before impact in complex, real-world scenes.

****

An Improved Scheme for Detection and Labelling in Johansson Displays

Claudio Fanti, Marzia Polito, Pietro Perona

Consider a number of moving points, where each point is attached to a joint of t he human body and projected onto an image plane. Johannson showed that humans ca n e■ortlessly detect and recog- nize the presence of other humans from such disp lays. This is true even when some of the body points are missing (e.g. because o f occlusion) and unrelated clutter points are added to the display. We are inter ested in replicating this ability in a machine. To this end, we present a labell ing and detection scheme in a probabilistic framework. Our method is based on re presenting the joint prob- ability density of positions and velocities of body p oints with a graphical model, and using Loopy Belief Propagation to calculate a likely interpretation of the scene. Furthermore, we introduce a global variable representing the body’s centroid. Experiments on one motion-captured sequence su ggest that our scheme improves on the accuracy of a previous approach based on t riangulated graph- ical models, especially when very few parts are visible. The im- provement is due both to the more general graph structure we use and, more s igni■cantly, to the introduction of the centroid variable.

****

1-norm Support Vector Machines

Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor Hastie

The standard 2-norm SVM is known for its good performance in two- In this paper, we consider the 1-norm SVM. We class classi£cation. argue that the 1-norm SVM m ay have some advantage over the standard 2-norm SVM, especially when there are r edundant noise features. We also propose an ef£cient algorithm that computes the whole solution path of the 1-norm SVM, hence facilitates adaptive selection of the tuning parameter for the 1-norm SVM.

****
Envelope-based Planning in Relational MDPs
Natalia Gardiol, Leslie Kaelbling
A mobile robot acting in the world is faced with a large amount of sen- sory dat a and uncertainty in its action outcomes. Indeed, almost all in- teresting seque ntial decision-making domains involve large state spaces and large, stochastic a ction sets. We investigate a way to act intelli- gently as quickly as possible i n domains where ■nding a complete policy would take a hopelessly long time. This approach, Relational Envelope- based Planning (REBP) tackles large, noisy probl ems along two axes. First, describing a domain as a relational MDP (instead of a s an atomic or propositionally-factored MDP) allows problem structure and dynam- ics to be captured compactly with a small set of probabilistic, relational rule s. Second, an envelope-based approach to planning lets an agent be- gin acting q uickly within a restricted part of the full state space and to judiciously expan d its envelope as resources permit.
****
Necessary Intransitive Likelihood-Ratio Classifiers
Gang Ji, Jeff A. Bilmes
In pattern classi■cation tasks, errors are introduced because of differ- ences b etween the true model and the one obtained via model estimation. Using likelihoo d-ratio based classi■cation, it is possible to correct for this discrepancy by ■ nding class-pair speci■c terms to adjust the likelihood ratio directly, and that can make class-pair preference relationships in- transitive. In this work, we i ntroduce new methodology that makes nec- essary corrections to the likelihood ra tio, speci■cally those that are nec- essary to achieve perfect classi■cation (bu t not perfect likelihood-ratio correction which can be overkill). The new correc tions, while weaker than previously reported such adjustments, are analytically challenging since they involve discontinuous functions, therefore requiring seve ral approximations. We test a number of these new schemes on an isolated- word s peech recognition task as well as on the UCI machine learning data sets. Results show that by using the bias terms calculated in this new way, classi■cation acc uracy can substantially improve over both the baseline and over our previous res ults.
****
Simplicial Mixtures of Markov Chains: Distributed Modelling of Dynamic User Prof iles
Mark Girolami, Ata Kabán
To provide a compact generative representation of the sequential activ- ity of a number of individuals within a group there is a tradeoff between the de■nition of individual speci■c and global models. This paper pro- poses a linear-time dis tributed model for ■nite state symbolic sequences representing traces of individ ual user activity by making the assump- tion that heterogeneous user behavior ma y be ‘explained’ by a relatively small number of common structurally simple beha vioral patterns which may interleave randomly in a user-speci■c proportion. The results of an empirical study on three different sources of user traces indicate s that this modelling approach provides an ef■cient representation scheme, re- ■ ected by improved prediction performance as well as providing low- complexity an d intuitively interpretable representations.
****
Bounded Finite State Controllers
Pascal Poupart, Craig Boutilier
We describe a new approximation algorithm for solving partially observ- able MDP s. Our bounded policy iteration approach searches through the space of bounded-s ize, stochastic ■nite state controllers, combining sev- eral advantages of gradi ent ascent (ef■ciency, search through restricted controller space) and policy it eration (less vulnerability to local optima).
****
Sparseness of Support Vector Machines---Some Asymptotically Sharp Bounds
Ingo Steinwart
The decision functions constructed by support vector machines (SVM’s) usually de

pend only on a subset of the training set—the so-called support vectors. We deri ve asymptotically sharp lower and upper bounds on the number of support vectors for several standard types of SVM's. In par- ticular, we show for the Gaussian R BF kernel that the fraction of support vectors tends to twice the Bayes risk for the L1-SVM, to the probability of noise for the L2-SVM, and to 1 for the LS-SVM .

****

Towards Social Robots: Automatic Evaluation of Human-Robot Interaction by Facial Expression Classification

G.C. Littlewort, M.S. Bartlett, I.R. Fasel, J. Chenu, T. Kanda, H. Ishiguro, J.R . Movellan

Computer animated agents and robots bring a social dimension to hu- man computer interaction and force us to think in new ways about how computers could be used in daily life. Face to face communication is a real-time process operating at a time scale of less than a second. In this paper we present progress on a percep tual primitive to automatically detect frontal faces in the video stream and cod e them with respect to 7 dimensions in real time: neutral, anger, disgust, fear, joy, sadness, sur- prise. The face ■nder employs a cascade of feature detectors trained with boosting techniques [13, 2]. The expression recognizer employs a n ovel combination of Adaboost and SVM's. The generalization performance to new su bjects for a 7-way forced choice was 93.3% and 97% correct on two publicly avail able datasets. The outputs of the classi■er change smoothly as a function of tim e, providing a potentially valuable repre- sentation to code facial expression d ynamics in a fully automatic and unobtrusive manner. The system was deployed and evaluated for mea- suring spontaneous facial expressions in the ■eld in an appl ication for automatic assessment of human-robot interaction.

****

Entrainment of Silicon Central Pattern Generators for Legged Locomotory Control

Francesco Tenore, Ralph Etienne-Cummings, M. Lewis

We have constructed a second generation CPG chip capable of generating the neces sary timing to control the leg of a walking machine. We demonstrate improvements over a previous chip by moving toward a significantly mor e versatile device. This includes a larger number of silicon neurons, more sophisticated neurons including voltage dependent charging and rel ative and absolute refractory periods, and enhanced programmability of ne ural networks. This chip builds on the basic results achieved on a previous ch ip and expands its versatility to get closer to a self-contained loco motion controller for walking robots. 1

****

Fast Embedding of Sparse Similarity Graphs

John Platt

This paper applies fast sparse multidimensional scaling (MDS) to a large graph o f music similarity, with 267K vertices that represent artists, al- bums, and tra cks; and 3.22M edges that represent similarity between those entities. Once vert ices are assigned locations in a Euclidean space, the locations can be used to b rowse music and to generate playlists. MDS on very large sparse graphs can be ef fectively performed by a family of algorithms called Rectangular Dijsktra (RD) M DS algorithms. These RD algorithms operate on a dense rectangular slice of the d istance matrix, created by calling Dijsktra a constant number of times. Two RD a lgorithms are compared: Landmark MDS, which uses the Nystr(cid:246)m ap- proxima tion to perform MDS; and a new algorithm called Fast Sparse Embedding, which use s FastMap. These algorithms compare favorably to Laplacian Eigenmaps, both in te rms of speed and embedding quality.

****

Online Passive-Aggressive Algorithms

Shai Shalev-shwartz, Koby Crammer, Ofer Dekel, Yoram Singer

We present a uni■ed view for online classi■cation, regression, and uni- class pr oblems. This view leads to a single algorithmic framework for the three problems . We prove worst case loss bounds for various algorithms for both the realizable case and the non-realizable case. A conversion of our main online algorithm to

the setting of batch learning is also dis- cussed. The end result is new algorit hms and accompanying loss bounds for the hinge-loss.
****

Training a Quantum Neural Network
Bob Ricks, Dan Ventura
Most proposals for quantum neural networks have skipped over the prob- lem of ho w to train the networks. The mechanics of quantum computing are different enough from classical computing that the issue of training should be treated in detail . We propose a simple quantum neural network and a training method for it. It ca n be shown that this algorithm works in quantum systems. Results on several real -world data sets show that this algorithm can train the proposed quantum neural networks, and that it has some advantages over classical learning algorithms.
****

Perspectives on Sparse Bayesian Learning
Jason Palmer, Bhaskar Rao, David Wipf
Recently, relevance vector machines (RVM) have been fashioned from a sparse Baye sian learning (SBL) framework to perform supervised learn- ing using a weight pr ior that encourages sparsity of representation. The methodology incorporates an additional set of hyperparameters govern- ing the prior, one for each weight, an d then adopts a speci■c approxi- mation to the full marginalization over all wei ghts and hyperparameters. Despite its empirical success however, no rigorous mot ivation for this particular approximation is currently available. To address thi s issue, we demonstrate that SBL can be recast as the application of a rigorous vari- ational approximation to the full model by expressing the prior in a dual form. This formulation obviates the necessity of assuming any hyperpri- ors and leads to natural, intuitive explanations of why sparsity is achieved in practice .
****

One Microphone Blind Dereverberation Based on Quasi-periodicity of Speech Signal s
Tomohiro Nakatani, Masato Miyoshi, Keisuke Kinoshita
Speech dereverberation is desirable with a view to achieving, for exam- ple, rob ust speech recognition in the real world. However, it is still a chal- lenging p roblem, especially when using a single microphone. Although blind equalization t echniques have been exploited, they cannot deal with speech signals appropriatel y because their assumptions are not satis■ed by speech signals. We propose a new dereverberation principle based on an inherent property of speech signals, name ly quasi-periodicity. The present methods learn the dereverberation ■lter from a lot of speech data with no prior knowledge of the data, and can achieve high qu ality speech dereverberation especially when the reverberation time is long.
****

Multiple Instance Learning via Disjunctive Programming Boosting
Stuart Andrews, Thomas Hofmann
Learning from ambiguous training data is highly relevant in many applications. W e present a new learning algorithm for classi■cation problems where labels are a ssociated with sets of pattern instead of individual patterns. This encompasses multiple instance learn- ing as a special case. Our approach is based on a gener alization of linear programming boosting and uses results from disjunctive progr amming to generate successively stronger linear relaxations of a discrete non-co nvex problem.
****

An Autonomous Robotic System for Mapping Abandoned Mines
David Ferguson, Aaron Morris, Dirk Hähnel, Christopher Baker, Zachary Omohundro, Carlos Reverte, Scott Thayer, Charles Whittaker, William Whittaker, Wolfram Bur gard, Sebastian Thrun
We present the software architecture of a robotic system for mapping abandoned m ines. The software is capable of acquiring consistent 2D maps of large mines wit h many cycles, represented as Markov random £elds. 3D C-space maps are acquired from local 3D range scans, which are used to identify navigable paths using A* s earch. Our system has been deployed in three abandoned mines, two of which inacc

essible to people, where it has acquired maps of unprecedented detail and accura cy.
****

Model Uncertainty in Classical Conditioning
Aaron C. Courville, Geoffrey J. Gordon, David Touretzky, Nathaniel Daw
We develop a framework based on Bayesian model averaging to explain how animals cope with uncertainty about contingencies in classical con- ditioning experiment s. Traditional accounts of conditioning ■t parame- ters within a ■xed generative model of reinforcer delivery; uncertainty over the model structure is not consi dered. We apply the theory to ex- plain the puzzling relationship between second -order conditioning and conditioned inhibition, two similar conditioning regimes that nonethe- less result in strongly divergent behavioral outcomes. According to the theory, second-order conditioning results when limited experience leads a nimals to prefer a simpler world model that produces spurious corre- lations; co nditioned inhibition results when a more complex model is justi■ed by additional experience.
****

Local Phase Coherence and the Perception of Blur
Zhou Wang, Eero Simoncelli
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.
****

New Algorithms for Efficient High Dimensional Non-parametric Classification
Ting liu, Andrew Moore, Alexander Gray
Alexander Gray
****

Large Margin Classifiers: Convex Loss, Low Noise, and Convergence Rates
Peter Bartlett, Michael Jordan, Jon Mcauliffe
Many classi■cation algorithms, including the support vector machine, boosting an d logistic regression, can be viewed as minimum contrast methods that minimize a convex surrogate of the 0-1 loss function. We characterize the statistical cons equences of using such a surrogate by pro- viding a general quantitative relatio nship between the risk as assessed us- ing the 0-1 loss and the risk as assessed using any nonnegative surrogate loss function. We show that this relationship g ives nontrivial bounds un- der the weakest possible condition on the loss functi on—that it satisfy a pointwise form of Fisher consistency for classi■cation. The relationship is based on a variational transformation of the loss function that is easy to compute in many applications. We also present a re■ned version of th is result in the case of low noise. Finally, we present applications of our resu lts to the estimation of convergence rates in the general setting of function cl asses that are scaled hulls of a ■nite-dimensional base class.
****

Design of Experiments via Information Theory
Liam Paninski
We discuss an idea for collecting data in a relatively ef■cient manner. Our poin t of view is Bayesian and information-theoretic: on any given trial, we want to adaptively choose the input in such a way that the mutual in- formation between the (unknown) state of the system and the (stochastic) output is maximal, given any prior information (including data collected on any previous trials). We prov e a theorem that quanti■es the effective- ness of this strategy and give a few i llustrative examples comparing the performance of this adaptive technique to tha t of the more usual nonadap- tive experimental design. For example, we are able to explicitly calculate the asymptotic relative ef■ciency of the “staircase meth od” widely em- ployed in psychophysics research, and to demonstrate the dependen ce of this ef■ciency on the form of the “psychometric function” underlying the o utput responses.
****

Minimax Embeddings

Matthew Brand
Spectral methods for nonlinear dimensionality reduction (NLDR) impose a neighbor hood graph on point data and compute eigenfunctions of a quadratic form generate d from the graph. We introduce a more general and more robust formulation of NLD R based on the singular value de- composition (SVD). In this framework, most spe ctral NLDR principles can be recovered by taking a subset of the constraints in a quadratic form built from local nullspaces on the manifold. The minimax formul ation also opens up an interesting class of methods in which the graph is “dec- orated” with information at the vertices, offering discrete or continuous maps, reduced computational complexity, and immunity to some solu- tion instabilities of eigenfunction approaches. Apropos, we show almost all NLDR methods based on e igenvalue decompositions (EVD) have a so- lution instability that increases fast er than problem size. This pathology can be observed (and corrected via the mini max formulation) in problems as small as N < 100 points.
****

Efficient and Robust Feature Extraction by Maximum Margin Criterion
Haifeng Li, Tao Jiang, Keshu Zhang
A new feature extraction criterion, maximum margin criterion (MMC), is proposed in this paper. This new criterion is general in the sense that, when combined wi th a suitable constraint, it can actually give rise to the most popular feature extractor in the literature, linear discriminate analysis (LDA). We derive a new feature extractor based on MMC using a different constraint that does not depen d on the nonsingularity of the within-class scatter matrix Sw. Such a dependence is a major drawback of LDA especially when the sample size is small. The kernel ized (nonlin- ear) counterpart of this linear feature extractor is also establis hed in this paper. Our preliminary experimental results on face images demonstra te that the new feature extractors are ef■cient and stable.
****

Training fMRI Classifiers to Detect Cognitive States across Multiple Human Subje cts
Xuerui Wang, Rebecca Hutchinson, Tom M. Mitchell
We consider learning to classify cognitive states of human subjects, based on th eir brain activity observed via functional Magnetic Resonance Imaging (fMRI). Th is problem is important because such classi■ers con- stitute “virtual sensors” o f hidden cognitive states, which may be useful in cognitive science research and clinical applications. In recent work, Mitchell, et al. [6,7,9] have demonstrat ed the feasibility of training such classi■ers for individual human subjects (e. g., to distinguish whether the subject is reading an ambiguous or unambiguous se ntence, or whether they are reading a noun or a verb). Here we extend that line of research, exploring how to train classi■ers that can be applied across multip le hu- man subjects, including subjects who were not involved in training the cl assi■er. We describe the design of several machine learning approaches to traini ng multiple-subject classi■ers, and report experimental results demonstrating th e success of these methods in learning cross-subject classi■ers for two differen t fMRI data sets.
****

Applying Metric-Trees to Belief-Point POMDPs
Joelle Pineau, Geoffrey J. Gordon, Sebastian Thrun
Recent developments in grid-based and point-based approximation algo- rithms for POMDPs have greatly improved the tractability of POMDP planning. These approach es operate on sets of belief points by individ- ually learning a value function for each point. In reality, belief points exist in a highly-structured metric si mplex, but current POMDP algo- rithms do not exploit this property. This paper p resents a new metric-tree algorithm which can be used in the context of POMDP pl anning to sort belief points spatially, and then perform fast value function upd ates over groups of points. We present results showing that this approach can re - duce computation in point-based POMDP algorithms for a wide range of problems.
****

Nonlinear Filtering of Electron Micrographs by Means of Support Vector Regressio n

Roland Vollgraf, Michael Scholz, Ian Meinertzhagen, Klaus Obermayer
Nonlinear (cid:12)ltering can solve very complex problems, but typically involve very time consuming calculations. Here we show that for (cid:12)lters that are constructed as a RBF network with Gaussian basis functions, a decomposition into linear (cid:12)lters exists, which can be computed e(cid:14)ciently in the frequency domain, yielding dramatic improvement in speed. We present an application of this idea to image processing. In electron micrograph images of photoreceptor terminals of the fruit (cid:13)y, Drosophila, synaptic vesicles containing neurotransmitter should be detected and labeled automatically. We use hand labels, provided by human experts, to learn a RBF (cid:12)lter using Support Vector Regression with Gaussian kernels. We will show that the resulting nonlinear (cid:12)lter solves the task to a degree of accuracy, which is close to what can be achieved by human experts. This allows the very time consuming task of data evaluation to be done e(cid:14)ciently.
****

Application of SVMs for Colour Classification and Collision Detection with AIBO Robots
Michael Quinlan, Stephan Chalup, Richard Middleton
This article addresses the issues of colour classi■cation and collision de- tection as they occur in the legged league robot soccer environment of RoboCup. We show how the method of one-class classi■cation with sup- port vector machines (SVMs) can be applied to solve these tasks satisfac- torily using the limited hardware capacity of the prescribed Sony AIBO quadruped robots. The experimental evaluation shows an improvement over our previous methods of ellipse ■tting for colour classi■cation and the statistical approach used for collision detection.
****

Warped Gaussian Processes
Edward Snelson, Zoubin Ghahramani, Carl Rasmussen
We generalise the Gaussian process (GP) framework for regression by learning a nonlinear transformation of the GP outputs. This allows for non-Gaussian processes and non-Gaussian noise. The learning algo- rithm chooses a nonlinear transformation such that transformed data is well-modelled by a GP. This can be seen as including a preprocessing transformation as an integral part of the probabilistic modelling problem, rather than as an ad-hoc step. We demonstrate on several real regression problems that learning the transformation can lead to signi■cantly better performance than using a regular GP, or a GP with a ■xed transformation.
****

Online Learning via Global Feedback for Phrase Recognition
Xavier Carreras, Lluís Màrquez
This work presents an architecture based on perceptrons to recognize phrase structures, and an online learning algorithm to train the percep- trons together and dependently. The recognition strategy applies learning in two layers: a ■ltering layer, which reduces the search space by identi- fying plausible phrase candidates, and a ranking layer, which recursively builds the optimal phrase structure. We provide a recognition-based feed- back rule which re■ects to each local function its committed errors from a global point of view, and allows to train them together online as percep- trons. Experimentation on a syntactic parsing problem, the recognition of clause hierarchies, improves state-of-the-art results and evinces the advantages of our global training method over optimizing each function locally and independently.
****

Iterative Scaled Trust-Region Learning in Krylov Subspaces via Pearlmutter's Implicit Sparse Hessian
Eiji Mizutani, James Demmel
The online incremental gradient (or backpropagation) algorithm is widely considered to be the fastest method for solving large-scale neural-network (NN) learning problems. In contrast, we show that an appropriately implemented iterative batch-mode (or block-mode) learning method can be much faster. For example, it is three times faster in the UCI letter classi■cation problem (26 outputs, 16,000 data items, 6,066 parameters with a two-hidden-layer multilayer perceptron) and 35

3 times faster in a nonlinear regression problem arising in color recipe predict ion (10 outputs, 1,000 data items, 2,210 parameters with a neuro-fuzzy modular n etwork). The three principal innovative ingredients in our algorithm are the fol lowing: First, we use scaled trust-region regularization with inner-outer it- er ation to solve the associated “overdetermined” nonlinear least squares problem, where the inner iteration performs a truncated (or inexact) Newton method. Secon d, we employ Pearlmutter’s implicit sparse Hessian matrix-vector multiply algori thm to con- struct the Krylov subspaces used to solve for the truncated New- ton update. Third, we exploit sparsity (for preconditioning) in the matrices result ing from the NNs having many outputs.

****

AUC Optimization vs. Error Rate Minimization

Corinna Cortes, Mehryar Mohri

The area under an ROC curve (AUC) is a criterion used in many appli- cations to measure the quality of a classi■cation algorithm. However, the objective functio n optimized in most of these algorithms is the error rate and not the AUC value. We give a detailed statistical analysis of the relationship between the AUC and the error rate, including the ■rst exact expression of the expected value and t he variance of the AUC for a ■xed error rate. Our results show that the average AUC is monotonically in- creasing as a function of the classi■cation accuracy, b ut that the standard deviation for uneven distributions and higher error rates i s noticeable. Thus, algorithms designed to minimize the error rate may not lead to the best possible AUC values. We show that, under certain conditions, the glo bal function optimized by the RankBoost algorithm is exactly the AUC. We report the results of our experiments with RankBoost in several datasets demonstrating the bene■ts of an algorithm speci■cally designed to globally optimize the AUC ov er other existing algorithms optimizing an approximation of the AUC or only loca lly optimizing the AUC.

****

Semi-Supervised Learning with Trees

Charles Kemp, Thomas Griffiths, Sean Stromsten, Joshua Tenenbaum

We describe a nonparametric Bayesian approach to generalizing from few labeled e xamples, guided by a larger set of unlabeled objects and the assumption of a lat ent tree-structure to the domain. The tree (or a distribution over trees) may be inferred using the unlabeled data. A prior over concepts generated by a mutatio n process on the inferred tree(s) allows ef■cient computation of the optimal Bay esian classi■cation func- tion from the labeled examples. We test our approach o n eight real-world datasets.

****

Finding the M Most Probable Configurations using Loopy Belief Propagation

Chen Yanover, Yair Weiss

Loopy belief propagation (BP) has been successfully used in a num- ber of di–cul t graphical models to ■nd the most probable con■gu- ration of the hidden variabl es. In applications ranging from protein folding to image analysis one would lik e to ■nd not just the best con■guration but rather the top M . While this proble m has been solved using the junction tree formalism, in many real world prob- le ms the clique size in the junction tree is prohibitively large. In this work we address the problem of ■nding the M best con■gura- tions when exact inference is impossible. We start by developing a new exact inference algorithm for calculat - ing the best con■gurations that uses only max-marginals. For ap- proximate inf erence, we replace the max-marginals with the beliefs calculated using max-produ ct BP and generalized BP. We show em- pirically that the algorithm can accuratel y and rapidly approximate the M best con■gurations in graphs with hundreds of va riables.

****

A Nonlinear Predictive State Representation

Matthew Rudary, Satinder Singh

Predictive state representations (PSRs) use predictions of a set of tests to rep resent the state of controlled dynamical systems. One reason why this representa tion is exciting as an alternative to partially observable Markov decision proce

sses (POMDPs) is that PSR models of dynamical systems may be much more compact t han POMDP models. Empirical work on PSRs to date has focused on linear PSRs, whi ch have not allowed for compression relative to POMDPs. We introduce a new notio n of tests which allows us to de■ne a new type of PSR that is nonlinear in gener al and allows for exponential compression in some deterministic dynami- cal syst ems. These new tests, called e-tests, are related to the tests used by Rivest an d Schapire [1] in their work with the diversity representation, but our PSR avoi ds some of the pitfalls of their representation—in partic- ular, its potential t o be exponentially larger than the equivalent POMDP.

****

A Recurrent Model of Orientation Maps with Simple and Complex Cells
Paul Merolla, Kwabena A. Boahen
that utilizes

****

Predicting Speech Intelligibility from a Population of Neurons
Jeff Bondy, Ian Bruce, Suzanna Becker, Simon Haykin
Simon Haykin

****

Different Cortico-Basal Ganglia Loops Specialize in Reward Prediction at Differe nt Time Scales
Saori Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, Shigeto Yam awaki
To understand the brain mechanisms involved in reward prediction on different t ime scales, we developed a Markov decision task that requires prediction of bot h immediate and future rewards, and ana- lyzed subjects' brain activities using functional MRI. We estimated the time course of reward prediction and reward pr ediction error on different time scales from subjects' performance data, and us ed them as the explanatory variables for SPM analysis. We found topog- raphic maps of different time scales in medial frontal cortex and st riatum. The result suggests that different cortico-basal ganglia loops a re specialized for reward prediction on different time scales.

****

Measure Based Regularization
Olivier Bousquet, Olivier Chapelle, Matthias Hein
We address in this paper the question of how the knowledge of the marginal distr ibution P (x) can be incorporated in a learning algorithm. We suggest three theo retical methods for taking into account this distribution for regularization and provide links to existing graph-based semi-supervised learning algorithms. We a lso propose practical implementations.

****

Geometric Clustering Using the Information Bottleneck Method
Susanne Still, William Bialek, Léon Bottou
We argue that K–means and deterministic annealing algorithms for geo- metric clu stering can be derived from the more general Information Bot- tleneck approach. If we cluster the identities of data points to preserve information about their location, the set of optimal solutions is massively degenerate. But if we treat the equations that de■ne the optimal solution as an iterative algorithm, then a set of "smooth" initial conditions selects solutions with the desired geometrica l properties. In addition to concep- tual uni■cation, we argue that this approac h can be more ef■cient and robust than classic algorithms.

****

Sample Propagation
Mark Paskin
Rao–Blackwellization is an approximation technique for probabilistic in- ference that ■exibly combines exact inference with sampling. It is useful in models whe re conditioning on some of the variables leaves a sim- pler inference problem th at can be solved tractably. This paper presents Sample Propagation, an ef■cient implementation of Rao–Blackwellized approximate inference for a large class of m odels. Sample Propagation tightly integrates sampling with message passing in a junction tree, and is named for its simple, appealing structure: it walks the cl

usters of a junction tree, sampling some of the current cluster’s variables and then passing a message to one of its neighbors. We discuss the application of Sa mple Propagation to conditional Gaussian inference problems such as switching li near dynamical systems.
****

Gaussian Processes in Reinforcement Learning
Malte Kuss, Carl Rasmussen
We exploit some useful properties of Gaussian process (GP) regression models for reinforcement learning in continuous state spaces and dis- crete time. We demon strate how the GP model allows evaluation of the value function in closed form. The resulting policy iteration algorithm is demonstrated on a simple problem wit h a two dimensional state space. Further, we speculate that the intrinsic abilit y of GP models to charac- terise distributions of functions would allow the meth od to capture entire distributions over future values instead of merely their ex pectation, which has traditionally been the focus of much of reinforcement learn ing.
****

The Doubly Balanced Network of Spiking Neurons: A Memory Model with High Capacit y
Yuval Aviel, David Horn, Moshe Abeles
A balanced network leads to contradictory constraints on memory models, as exem plified in previous work on accommodation of synfire chains. Here we show that these constraints can be overcome by introducing a 'shadow' inhibitory pattern for each excitatory pattern of the model. This is interpreted as a double- bala nce principle, whereby there exists both global balance between average excitat ory and inhibitory currents and local balance between the currents carrying coh erent activity at any given time frame. This principle can be applied to networ ks with Hebbian cell assemblies, leading to a high capacity of the associative memory. The number of possible patterns is limited by a combinatorial constrai nt that turns out to be P=0.06N within the specific model that we employ. This limit is reached by the Hebbian cell assembly network. To the best of our knowl edge this is the first time that such high memory capacities are demonstrated in the asynchronous state of models of spiking neurons.
****

Hierarchical Topic Models and the Nested Chinese Restaurant Process
Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, David Blei
We address the problem of learning topic hierarchies from data. The model select ion problem in this domain is daunting—which of the large collection of possible trees to use? We take a Bayesian approach, gen- erating an appropriate prior vi a a distribution on partitions that we refer to as the nested Chinese restaurant process. This nonparametric prior al- lows arbitrarily large branching factors and readily accommodates grow- ing data collections. We build a hierarchical top ic model by combining this prior with a likelihood that is based on a hierarchic al variant of latent Dirichlet allocation. We illustrate our approach on simulat ed data and with an application to the modeling of NIPS abstracts.
****

A Sampled Texture Prior for Image Super-Resolution
Lyndsey Pickup, Stephen J. Roberts, Andrew Zisserman
Super-resolution aims to produce a high-resolution image from a set of one or mo re low-resolution images by recovering or inventing plausible high-frequency ima ge content. Typical approaches try to reconstruct a high-resolution image using the sub-pixel displacements of several low- resolution images, usually regulariz ed by a generic smoothness prior over the high-resolution image space. Other met hods use training data to learn low-to-high-resolution matches, and have been hi ghly successful even in the single-input-image case. Here we present a domain-sp eci■c im- age prior in the form of a p.d.f. based upon sampled images, and show that for certain types of super-resolution problems, this sample-based prior giv es a signi■cant improvement over other common multiple-image super-resolution te chniques.
****

Analytical Solution of Spike-timing Dependent Plasticity Based on Synaptic Biophysics

Bernd Porr, Ausra Saudargiene, Florentin Wörgötter

Spike timing plasticity (STDP) is a special form of synaptic plasticity where the relative timing of post- and presynaptic activity determines the change of the synaptic weight. On the postsynaptic side, active back- propagating spikes in dendrites seem to play a crucial role in the induc- tion of spike timing dependent plasticity. We argue that postsynaptically the temporal change of the membrane potential determines the weight change. Coming from the presynaptic side induction of STDP is closely related to the activation of NMDA channels. Therefore, we will calculate analytically the change of the synaptic weight by correlating the deriva- tive of the membrane potential with the activity of the NMDA channel. Thus, for this calculation we utilise biophysical variables of the physi- ologic al cell. The ■nal result shows a weight change curve which con- forms with measurements from biology. The positive part of the weight change curve is determined by the NMDA activation. The negative part of the weight change curve is determined by the membrane potential change. Therefore, the weight change curve should change its shape de- pending on the distance from the soma of the postsynaptic cell. We ■nd temporally asymmetric weight change close to the soma and temporally symmetric weight change in the distal dendrite.

****

Sparse Greedy Minimax Probability Machine Classification

Thomas R. Strohmann, Andrei Belitski, Gregory Grudic, Dennis DeCoste

The Minimax Probability Machine Classi■cation (MPMC) framework [Lanckriet et al., 2002] builds classi■ers by minimizing the maximum probability of misclassi■cation, and gives direct estimates of the proba- bilistic accuracy bound Ω. The only assumptions that MPMC makes is that good estimates of means and covariance matrices of the classes exist. However, as with Support Vector Machines, MPMC is computationally expensive and requires extensive cross validation experiments to choose kernels and kernel parameters that give good performance. In this paper we address the computational cost of MPMC by proposing an algorithm that constructs nonlinear sparse MPMC (SMPMC) models by incremen- tally adding basis functions (i.e. kernels) one at a time – greedily select- ing the next one that maximizes the accuracy bound Ω. SMPMC auto- matically chooses both kernel parameters and feature weights without us- ing computationally expensive cross validation. Therefore the SMPMC algorithm simultaneously addresses the problem of kernel selection and feature selection (i.e. feature weighting), based solely on maximizing the accuracy bound Ω. Experimental results indicate that we can obtain reli- able bounds Ω, as well as test set accuracies that are comparable to state of the art classi■cation algorithms.

****

Approximate Policy Iteration with a Policy Language Bias

Alan Fern, Sungwook Yoon, Robert Givan

We explore approximate policy iteration, replacing the usual cost- function learning step with a learning step in policy space. We give policy-language biases that enable solution of very large relational Markov decision processes (MDPs) that no previous technique can solve. In particular, we induce high-quality domain-speci■c planners for clas- sical planning domains (both deterministic and stochastic variants) by solving such domains as extremely large MDPs.

****

Information Bottleneck for Gaussian Variables

Gal Chechik, Amir Globerson, Naftali Tishby, Yair Weiss

The problem of extracting the relevant aspects of data was ad- dressed through the information bottleneck (IB) method, by (soft) clustering one variable while preserving information about another - relevance - variable. An interesting question addressed in the current work is the extension of these ideas to obtain continuous representations that preserve relevant information, rather than dis- crete clusters. We give a formal de■nition of the general continuous IB problem and obtain an analytic solution for the optimal repre- sentation for the important case of multivariate Gaussian variables. The obtained optimal representation is a

noisy linear projection to eigenvectors of the normalized correlation matrix §x jy§¡1 x , which is also the basis obtained in Canonical Correlation Analysis. Ho w- ever, in Gaussian IB, the compression tradeo■ parameter uniquely determines t he dimension, as well as the scale of each eigenvector. This introduces a novel interpretation where solutions of di■erent ranks lie on a continuum parametrized by the compression level. Our analysis also provides an analytic expression for the optimal tradeo■ - the information curve - in terms of the eigenvalue spec- trum.

****

Laplace Propagation

Eleazar Eskin, Alex Smola, S.v.n. Vishwanathan

We present a novel method for approximate inference in Bayesian mod- els and reg ularized risk functionals. It is based on the propagation of mean and variance d erived from the Laplace approximation of condi- tional probabilities in factoriz ing distributions, much akin to Minka’s Expectation Propagation. In the jointly normal case, it coincides with the latter and belief propagation, whereas in the general case, it provides an optimization strategy containing Support Vector ch unking, the Bayes Committee Machine, and Gaussian Process chunking as special ca ses.

****

Error Bounds for Transductive Learning via Compression and Clustering

Philip Derbeko, Ran El-Yaniv, Ron Meir

This paper is concerned with transductive learning. Although transduc- tion appe ars to be an easier task than induction, there have not been many provably usefu l algorithms and bounds for transduction. We present ex- plicit error bounds for transduction and derive a general technique for devising bounds within this set ting. The technique is applied to derive error bounds for compression schemes su ch as (transductive) SVMs and for transduction algorithms based on clustering.

****

Approximate Expectation Maximization

Tom Heskes, Onno Zoeter, Wim Wiegerinck

We discuss the integration of the expectation-maximization (EM) algorithm for ma ximum likelihood learning of Bayesian networks with belief propagation algorithm s for approximate inference. Specifically we propose to combine the outer-loop s tep of convergent belief propagation algorithms with the M-step of the EM algori thm. This then yields an approximate EM algorithm that is essentially still doub le loop, with the important advantage of an inner loop that is guaranteed to con verge. Simulations illustrate the merits of such an approach.

****

A Fast Multi-Resolution Method for Detection of Significant Spatial Disease Clus ters

Daniel Neill, Andrew Moore

Given an N(cid:2)N grid of squares, where each square has a count and an un- der lying population, our goal is to ■nd the square region with the highest density, and to calculate its signi■cance by randomization. Any density measure D, depen dent on the total count and total population of a re- gion, can be used. For exa mple, if each count represents the number of disease cases occurring in that squ are, we can use Kulldorff’s spatial scan statistic DK to ■nd the most signi■cant spatial disease cluster. A naive approach to ■nding the maximum density region requires O(N 3) time, and is generally computationally infeasible. We present a novel algorithm which partitions the grid into overlapping regions, bounds the m aximum score of subregions contained in each region, and prunes re- gions which cannot contain the maximum density region. For suf■ciently dense regions, this m ethod ■nds the maximum density region in optimal O(N2) time, in practice resulti ng in signi■cant (10-200x) speedups.

****

Computing Gaussian Mixture Models with EM Using Equivalence Constraints

Noam Shental, Aharon Bar-hillel, Tomer Hertz, Daphna Weinshall

Density estimation with Gaussian Mixture Models is a popular gener- ative techni que used also for clustering. We develop a framework to incorporate side informa

tion in the form of equivalence constraints into the model estimation procedure. Equivalence constraints are de■ned on pairs of data points, indicating whether the points arise from the same source (positive constraints) or from different s ources (negative con- straints). Such constraints can be gathered automatically in some learn- ing problems, and are a natural form of supervision in others. Fo r the estimation of model parameters we present a closed form EM procedure which handles positive constraints, and a Generalized EM procedure us- ing a Markov n et which handles negative constraints. Using publicly available data sets we dem onstrate that such side information can lead to considerable improvement in clus tering tasks, and that our algorithm is preferable to two other suggested method s using the same type of side information.

****

Convex Methods for Transduction

Tijl Bie, Nello Cristianini

The 2-class transduction problem, as formulated by Vapnik [1], involves ■nding a separating hyperplane for a labelled data set that is also maximally distant fr om a given set of unlabelled test points. In this form, the problem has exponent ial computational complexity in the size of the working set. So far it has been attacked by means of integer programming techniques [2] that do not scale to rea sonable problem sizes, or by local search procedures [3]. In this paper we prese nt a relaxation of this task based on semi- de■nite programming (SDP), resulting in a convex optimization problem that has polynomial complexity in the size of the data set. The results are very encouraging for mid sized data sets, however the cost is still too high for large scale problems, due to the high di- mension al search space. To this end, we restrict the feasible region by introducing an approximation based on solving an eigenproblem. With this approximation, the com putational cost of the algorithm is such that problems with more than 1000 point s can be treated.

****

Kernel Dimensionality Reduction for Supervised Learning

Kenji Fukumizu, Francis Bach, Michael Jordan

We propose a novel method of dimensionality reduction for supervised learning. G iven a regression or classi■cation problem in which we wish to predict a variabl e Y from an explanatory vector X, we treat the prob- lem of dimensionality reduc tion as that of ■nding a low-dimensional “ef- fective subspace” of X which retai ns the statistical relationship between X and Y . We show that this problem can be formulated in terms of conditional independence. To turn this formulation int o an optimization problem, we characterize the notion of conditional independenc e using covariance operators on reproducing kernel Hilbert spaces; this allows u s to derive a contrast function for estimation of the effective subspace. Un- li ke many conventional methods, the proposed method requires neither assumptions o n the marginal distribution of X, nor a parametric model of the conditional dist ribution of Y .

****

Learning with Local and Global Consistency

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, Bernhard Schölkopf

We consider the general problem of learning from labeled and unlabeled data, whi ch is often called semi-supervised learning or transductive in- ference. A princ ipled approach to semi-supervised learning is to design a classifying function w hich is suf(cid:2)ciently smooth with respect to the intrinsic structure collect ively revealed by known labeled and unlabeled points. We present a simple algori thm to obtain such a smooth solution. Our method yields encouraging experimental results on a number of clas- si(cid:2)cation problems and demonstrates effectiv e use of unlabeled data.

****

Max-Margin Markov Networks

Ben Taskar, Carlos Guestrin, Daphne Koller

In typical classi■cation tasks, we seek a function which assigns a label to a si n- gle object. Kernel-based approaches, such as support vector machines (SVMs), which maximize the margin of con■dence of the classi■er, are the method of choic

e for many such tasks. Their popularity stems both from the ability to use high-dimensional feature spaces, and from their strong theoretical guaran- tees. However, many real-world tasks involve sequential, spatial, or structured data, where multiple labels must be assigned. Existing kernel-based methods ig- nore structure in the problem, assigning labels independently to each object, los- ing much useful information. Conversely, probabilistic graphical models, such as Markov networks, can represent correlations between labels, by exploiting problem structure, but cannot handle high-dimensional feature spaces, and lack strong theoretical generalization guarantees. In this paper, we present a new framework that combines the advantages of both approaches: Maximum mar- gin Markov (M3) networks incorporate both kernels, which ef■ciently deal with high-dimensional features, and the ability to capture correlations in structured data. We present an ef■cient algorithm for learning M3 networks based on a compact quadratic program formulation. We provide a new theoretical bound for generalization in structured domains. Experiments on the task of handwrit- ten character recognition and collective hypertext classi■cation demonstrate very signi■cant gains over previous approaches.

****

Factorization with Uncertainty and Missing Data: Exploiting Temporal Coherence

Amit Gruber, Yair Weiss

The problem of \Structure From Motion" is a central problem in vision: given the 2D locations of certain points we wish to recover the camera motion and the 3D coordinates of the points. Un- der simpli■ed camera models, the problem reduces to factorizing a measurement matrix into the product of two low rank matrices. Each element of the measurement matrix contains the position of a point in a particular image. When all elements are observed, the problem can be solved trivially using SVD, but in any realistic sit- uation many elements of the matrix are missing and the ones that are observed have a di■erent directional uncertainty. Under these conditions, most existing factorization algorithms fail while human perception is relatively unchanged. In this paper we use the well known EM algorithm for factor analy- sis to perform factorization. This allows us to easily handle missing data and measurement uncertainty and more importantly allows us to place a prior on the temporal trajectory of the latent variables (the camera position). We show that incorporating this prior gives a signi■cant improvement in performance in challenging image se- quences.

****

Sequential Bayesian Kernel Regression

Jaco Vermaak, Simon Godsill, Arnaud Doucet

We propose a method for sequential Bayesian kernel regression. As is the case for the popular Relevance Vector Machine (RVM) [10, 11], the method automatically identi■es the number and locations of the kernels. Our algorithm overcomes some of the computational dif■culties related to batch methods for kernel regression. It is non-iterative, and requires only a single pass over the data. It is thus applicable to truly sequen- tial data sets and batch data sets alike. The algorithm is based on a generalisation of Importance Sampling, which allows the design of in- tuitively simple and ef■cient proposal distributions for the model param- eters. Comparative results on two standard data sets show our algorithm to compare favourably with existing batch estimation strategies.

****

PAC-Bayesian Generic Chaining

Jean-yves Audibert, Olivier Bousquet

There exist many different generalization error bounds for classi■cation. Each of these bounds contains an improvement over the others for cer- tain situations. Our goal is to combine these different improvements into a single bound. In particular we combine the PAC-Bayes approach intro- duced by McAllester [1], which is interesting for averaging classi■ers, with the optimal union bound provided by the generic chaining technique developed by Fernique and Talagrand [2]. This combination is quite nat- ural since the generic chaining is based on the notion of majorizing mea- sures, which can be considered as priors on the set of classi■ers, and such priors also arise in the PAC-bayesian setting.

****
Ambiguous Model Learning Made Unambiguous with 1/f Priors
Gurinder Atwal, William Bialek
What happens to the optimal interpretation of noisy data when there exists more than one equally plausible interpretation of the data? In a Bayesian model-learning framework the answer depends on the prior ex- pectations of the dynamics of the model parameter that is to be inferred from the data. Local time constraints on the priors are insuf■cient to pick one interpretation over another. On the other hand, nonlocal time constraints, induced by a 1/f noise spectrum of the priors, is shown to permit learning of a speci■c model parameter even when there are in- ■nitely many equally plausible interpretations of the data. This transition is inferred by a remarkable mapping of the model estimation problem to a dissipative physical system, allowing the use of powerful statisti- cal mechanical methods to uncover the transition from indeterminate to determinate model learning.
****
Generalised Propagation for Fast Fourier Transforms with Partial or Missing Data
Amos J. Storkey
Discrete Fourier transforms and other related Fourier methods have been practically implementable due to the fast Fourier transform (FFT). However there are many situations where doing fast Fourier transforms without complete data would be desirable. In this pa- per it is recognised that formulating the FFT algorithm as a belief network allows suitable priors to be set for the Fourier coe(cid:14)cients. Furthermore e(cid:14)cient generalised belief propagation methods be- tween clusters of four nodes enable the Fourier coe(cid:14)cients to be inferred and the missing data to be estimated in near to O(n log n) time, where n is the total of the given and missing data points. This method is compared with a number of common approaches such as setting missing data to zero or to interpolation. It is tested on generated data and for a Fourier analysis of a damaged audio signal.
****
Can We Learn to Beat the Best Stock
Allan Borodin, Ran El-Yaniv, Vincent Gogan
A novel algorithm for actively trading stocks is presented. While tradi- tional universal algorithms (and technical trading heuristics) attempt to predict winners or trends, our approach relies on predictable statistical relations between all pairs of stocks in the market. Our empirical results on historical markets provide strong evidence that this type of techni- cal trading can “beat the market” and moreover, can beat the best stock in the market. In doing so we utilize a new idea for smoothing critical parameters in the context of expert learning.
****
An MCMC-Based Method of Comparing Connectionist Models in Cognitive Science
Woojae Kim, Daniel Navarro, Mark Pitt, In Myung
Despite the popularity of connectionist models in cognitive science, their performance can often be di■cult to evaluate. Inspired by the geometric approach to statistical model selection, we introduce a conceptually similar method to examine the global behavior of a connectionist model, by counting the number and types of response patterns it can simulate. The Markov Chain Monte Carlo-based algorithm that we constructed (cid:222)nds these patterns e■ciently. We demonstrate the approach using two localist network models of speech perception.
****
Dynamical Modeling with Kernels for Nonlinear Time Series Prediction
Liva Ralaivola, Florence d'Alché-Buc
We consider the question of predicting nonlinear time series. Kernel Dy- namical Modeling (KDM), a new method based on kernels, is proposed as an extension to linear dynamical models. The kernel trick is used twice: ■rst, to learn the parameters of the model, and second, to compute preimages of the time series predicted in the feature space by means of Support Vector Regression. Our model shows strong connection with the classic Kalman Filter model, with the kernel feature space as hidden state space. Kernel Dynamical Modeling is tested against two bench

mark time series and achieves high quality predictions.
****

Learning Non-Rigid 3D Shape from 2D Motion
Lorenzo Torresani, Aaron Hertzmann, Christoph Bregler
This paper presents an algorithm for learning the time-varying shape of a non-rigid 3D object from uncalibrated 2D tracking data. We model shape motion as a rigid component (rotation and translation) combined with a non-rigid deformation. Reconstruction is ill-posed if arbitrary deforma- tions are allowed. We constrain the problem by assuming that the object shape at each time instant is drawn from a Gaussian distribution. Based on this assumption, the algorithm simultaneously estimates 3D shape and motion for each time frame, learns the parameters of the Gaussian, and robustly ■lls-in missing data points. We then extend the algorithm to model temporal smoothness in object shape, thus allowing it to handle severe cases of missing data.
****

Discriminative Fields for Modeling Spatial Dependencies in Natural Images
Sanjiv Kumar, Martial Hebert
In this paper we present Discriminative Random Fields (DRF), a discrim- inative framework for the classi■cation of natural image regions by incor- porating neighborhood spatial dependencies in the labels as well as the observed data. The proposed model exploits local discriminative models and allows to relax the assumption of conditional independence of the observed data given the labels, commonly used in the Markov Random Field (MRF) framework. The parameters of the DRF model are learned using penalized maximum pseudo-likelihood method. Furthermore, the form of the DRF model allows the MAP inference for binary classi■ca- tion problems using the graph min-cut algorithms. The performance of the model was veri■ed on the synthetic as well as the real-world images. The DRF model outperforms the MRF model in the experiments.
****

Fast Algorithms for Large-State-Space HMMs with Applications to Web Usage Analysis
Pedro Felzenszwalb, Daniel Huttenlocher, Jon Kleinberg
In applying Hidden Markov Models to the analysis of massive data streams, it is often necessary to use an arti(cid:12)cially reduced set of states; this is due in large part to the fact that the basic HMM estimation algorithms have a quadratic dependence on the size of the state set. We present algorithms that reduce this computational bottleneck to linear or near-linear time, when the states can be embedded in an underlying grid of parameters. This type of state representation arises in many domains; in particular, we show an application to tra(cid:14)c analysis at a high-volume Web site.
****

A Probabilistic Model of Auditory Space Representation in the Barn Owl
Brian Fischer, Charles Anderson
The barn owl is a nocturnal hunter, capable of capturing prey using au- ditory information alone [1]. The neural basis for this localization be- havior is the existence of auditory neurons with spatial receptive ■elds [2]. We provide a mathematical description of the operations performed on auditory input signals by the barn owl that facilitate the creation of a representation of auditory space. To develop our model, we ■rst formu- late the sound localization problem solved by the barn owl as a statistical estimation problem. The implementation of the solution is constrained by the known neurobiology.
****

Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data
Neil Lawrence
In this paper we introduce a new underlying probabilistic model for prin- cipal component analysis (PCA). Our formulation interprets PCA as a particular Gaussian process prior on a mapping from a latent space to the observed data-space. We show that if the prior’s covariance func- tion constrains the mappings to be linear the model is equivalent to PCA, we then extend the model by considering less

restrictive covariance func- tions which allow non-linear mappings. This more general Gaussian pro- cess latent variable model (GPLVM) is then evaluated as an approach to the visualisation of high dimensional data for three different data- sets. Additionally our non-linear algorithm can be further kernelised leading to 'twin kernel PCA' in which a mapping between feature spaces occurs.

****

Tree-structured Approximations by Expectation Propagation

Yuan Qi, Tom Minka

Approximation structure plays an important role in inference on loopy graphs. As a tractable structure, tree approximations have been utilized in the variational method of Ghahramani & Jordan (1997) and the se- quential projection method of Frey et al. (2000). However, belief propa- gation represents each factor of the graph with a product of single-node messages. In this paper, belief propagation is extended to represent fac- tors with tree approximations, by way of the expectation propagation framework. That is, each factor sends a "message" to all pairs of nodes in a tree structure. The result is more accurate inferences and more fre- quent convergence than ordinary belief propagation, at a lower cost than variational trees or double-loop algorithms.

****

Boosting versus Covering

Kohei Hatano, Manfred K. K. Warmuth

We investigate improvements of AdaBoost that can exploit the fact that the weak hypotheses are one-sided, i.e. either all its positive (or negative) predictions are correct. In particular, for any set of m labeled examples consistent with a disjunction of k literals (which are one-sided in this case), AdaBoost constructs a consistent hypothesis by using O(k2 log m) iterations. On the other hand, a greedy set covering algorithm ■nds a consistent hypothesis of size O(k log m). Our primary question is whether there is a simple boosting algorithm that performs as well as the greedy set covering. We ■rst show that InfoBoost, a modi■catio n of AdaBoost pro- posed by Aslam for a di■erent purpose, does perform as well as the greedy set covering algorithm. We then show that AdaBoost requires Ω(k2 log m) iterations for learning k-literal disjunctions. We achieve this with an adversary construction and as well as in simple experiments based on arti■cial data . Further we give a vari- ant called SemiBoost that can handle the degenerate case when the given examples all have the same label. We conclude by showing that SemiBoost can be used to produce small conjunctions as well.

****

Learning Near-Pareto-Optimal Conventions in Polynomial Time

Xiaofeng Wang, Tuomas Sandholm

We study how to learn to play a Pareto-optimal strict Nash equilibrium when there exist multiple equilibria and agents may have different pref- erences among the equilibria. We focus on repeated coordination games of non-identical interest where agents do not know the game structure up front and receive noisy payoffs. We design ef■cient near-optimal al- gorithms for both the perfect monitoring and the imperfect monitoring setting(where the agents only observe their own payoffs and the joint actions).

****

Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes

Kevin P. Murphy, Antonio Torralba, William Freeman

Standard approaches to object detection focus on local patches of the image, and try to classify them as background or not. We propose to use the scene context (image as a whole) as an extra source of (global) information, to help resolve local ambiguities. We present a conditional random ■eld for jointly solving the tasks of object detection and scene classi■cation.

****

Approximability of Probability Distributions

Alina Beygelzimer, Irina Rish

We consider the question of how well a given distribution can be approx- imated with probabilistic graphical models. We introduce a new param- eter, effective t

reewidth, that captures the degree of approximability as a tradeoff between the accuracy and the complexity of approximation. We present a simple approach to an alyzing achievable tradeoffs that ex- ploits the threshold behavior of monotone graph properties, and provide experimental results that support the approach.
****

Gene Expression Clustering with Functional Mixture Models
Darya Chudova, Christopher Hart, Eric Mjolsness, Padhraic Smyth
We propose a functional mixture model for simultaneous clustering and alignment of sets of curves measured on a discrete time grid. The model is speci■cally tai lored to gene expression time course data. Each func- tional cluster center is a nonlinear combination of solutions of a simple linear differential equation tha t describes the change of individual mRNA levels when the synthesis and decay ra tes are constant. The mixture of continuous time parametric functional forms all ows one to (a) account for the heterogeneity in the observed pro■les, (b) align the pro■les in time by estimating real-valued time shifts, (c) capture the synth esis and decay of mRNA in the course of an experiment, and (d) regularize noisy pro■les by enforcing smoothness in the mean curves. We derive an EM algo- rithm for estimating the parameters of the model, and apply the proposed approach to t he set of cycling genes in yeast. The experiments show consistent improvement in predictive power and within cluster variance compared to regular Gaussian mixtu res.
****

Large Scale Online Learning
Léon Bottou, Yann Cun
We consider situations where training data is abundant and computing resources a re comparatively scarce. We argue that suitably designed on- line learning algor ithms asymptotically outperform any batch learning algorithm. Both theoretical a nd experimental evidences are presented.
****

Plasticity Kernels and Temporal Statistics
Peter Dayan, Michael Häusser, Michael London
Computational mysteries surround the kernels relating the magnitude and sign of changes in efficacy as a function of the time difference betwe en pre- and post-synaptic activity at a synapse. One important idea34 is tha t kernels result from fil(cid:173) tering, ie an attempt by synapses to eliminat e noise corrupting learning. This idea has hitherto been applied to trace lear ning rules; we apply it to experimentally-defined kernels, using it to reverse -engineer assumed signal statistics. We also extend it to consider the addition al goal for filtering of weighting learning according to statistical surpri se, as in the Z-score transform. This provides a fresh view of observed ker nels and can lead to different, and more natural, signal statistics.
****

Phonetic Speaker Recognition with Support Vector Machines
William Campbell, Joseph Campbell, Douglas Reynolds, Douglas Jones, Timothy Leek
A recent area of signi■cant progress in speaker recognition is the use of high l evel features—idiolect, phonetic relations, prosody, discourse structure, etc. A speaker not only has a distinctive acoustic sound but uses language in a charac teristic manner. Large corpora of speech data available in recent years allow ex perimentation with long term statistics of phone patterns, word patterns, etc. o f an individual. We propose the use of support vector machines and term frequenc y analysis of phone se- quences to model a given speaker. To this end, we explor e techniques for text categorization applied to the problem. We derive a new ker nel based upon a linearization of likelihood ratio scoring. We introduce a new p hone-based SVM speaker recognition approach that halves the er- ror rate of conv entional phone-based approaches.
****

Information Maximization in Noisy Channels : A Variational Approach
David Barber, Felix Agakov
The maximisation of information transmission over noisy channels is a common, al beit generally computationally di■cult problem. We approach the di■culty of comp

uting the mutual information for noisy channels by using a variational approxima tion. The re- sulting IM algorithm is analagous to the EM algorithm, yet max- im ises mutual information, as opposed to likelihood. We apply the method to severa l practical examples, including linear compression, population encoding and CDMA .

****

Variational Linear Response

Manfred Opper, Ole Winther

A general linear response method for deriving improved estimates of cor- relatio ns in the variational Bayes framework is presented. Three applica- tions are giv en and it is discussed how to use linear response as a general principle for imp roving mean ■eld approximations.

****

Circuit Optimization Predicts Dynamic Networks for Chemosensory Orientation in N ematode C. elegans

Nathan A. Dunn, John S. Conery, Shawn Lockery

The connectivity of the nervous system of the nematode Caenorhabdi- tis elegans has been described completely, but the analysis of the neu- ronal basis of behav ior in this system is just beginning. Here, we used an optimization algorithm to search for patterns of connectivity suf■- cient to compute the sensorimotor tra nsformation underlying C. elegans chemotaxis, a simple form of spatial orientati on behavior in which turn- ing probability is modulated by the rate of change of chemical concen- tration. Optimization produced differentiator networks with in hibitory feedback among all neurons. Further analysis showed that feedback reg- ulates the latency between sensory input and behavior. Common patterns of connec tivity between the model and biological networks suggest new functions for previ ously identi■ed connections in the C. elegans nervous system.

****

Learning to Find Pre-Images

Jason Weston, Bernhard Schölkopf, Gökhan Bakir

We consider the problem of reconstructing patterns from a feature map. Learning algorithms using kernels to operate in a reproducing kernel Hilbert space (RKHS) express their solutions in terms of input points mapped into the RKHS. We intro duce a technique based on kernel princi- pal component analysis and regression t o reconstruct corresponding pat- terns in the input space (aka pre-images) and r eview its performance in several applications requiring the construction of pre- images. The intro- duced technique avoids dif■cult and/or unstable numerical opt imization, is easy to implement and, unlike previous methods, permits the compu- tation of pre-images in discrete input spaces.

****

Sparse Representation and Its Applications in Blind Source Separation

Yuanqing Li, Shun-ichi Amari, Sergei Shishkin, Jianting Cao, Fanji Gu, Andrzej C ichocki

In this paper, sparse representation (factorization) of a data matrix is ■rst di scussed. An overcomplete basis matrix is estimated by using the K(cid:0)means me thod. We have proved that for the estimated overcom- plete basis matrix, the spa rse solution (coef■cient matrix) with minimum l1(cid:0)norm is unique with proba bility of one, which can be obtained using a linear programming algorithm. The c omparisons of the l1(cid:0)norm so- lution and the l0(cid:0)norm solution are al so presented, which can be used in recoverability analysis of blind source separ ation (BSS). Next, we ap- ply the sparse matrix factorization approach to BSS in the overcomplete case. Generally, if the sources are not suf■ciently sparse, we perform blind separation in the time-frequency domain after preprocessing the o bserved data using the wavelet packets transformation. Third, an EEG experimenta l data analysis example is presented to illustrate the useful- ness of the propo sed approach and demonstrate its performance. Two almost independent components obtained by the sparse representation method are selected for phase synchronizat ion analysis, and their peri- ods of signi■cant phase synchronization are found which are related to tasks. Finally, concluding remarks review the approach and state areas that require further study.

****
Probabilistic Inference in Human Sensorimotor Processing
Konrad Körding, Daniel M. Wolpert
When we learn a new motor skill, we have to contend with both the vari- ability inherent in our sensors and the task. The sensory uncertainty can be reduced by using information about the distribution of previously ex- perienced tasks. Here we impose a distribution on a novel sensorimotor task and manipulate the variability of the sensory feedback. We show that subjects internally represent both the distribution of the task as well as their sensory uncertainty. Moreover, they combine these two sources of information in a way that is qualitatively predicted by optimal Bayesian processing. We further analyze if the subjects can represent multimodal distributions such as mixtures of Gaussians. The results show that the CNS employs probabilistic models during sensorimotor learning even when the priors are multimodal.
****
From Algorithmic to Subjective Randomness
Thomas Griffiths, Joshua Tenenbaum
We explore the phenomena of subjective randomness as a case study in understanding how people discover structure embedded in noise. We present a rational account of randomness perception based on the statis- tical problem of model selection: given a stimulus, inferring whether the process that generated it was random or regular. Inspired by the mathe- matical definition of randomness given by Kolmogorov complexity, we characterize regularity in terms of a hierarchy of automata that augment a finite controller with different forms of memory. We find that the reg- ularities detected in binary sequences depend upon presentation format, and that the kinds of automata that can identify these regularities are in- formative about the cognitive processes engaged by different formats.
****
Eye Micro-movements Improve Stimulus Detection Beyond the Nyquist Limit in the Peripheral Retina
Matthias Hennig, Florentin Wörgötter
Even under perfect ■xation the human eye is under steady motion (tremor, microsaccades, slow drift). The “dynamic” theory of vi- sion [1, 2] states that eye-movements can improve hyperacuity. Accord- ing to this theory, eye movements are thought to create variable spatial excitation patterns on the photoreceptor grid, which will allow for better spatiotemporal summation at later stages. We reexamine this theory us- ing a realistic model of the vertebrate retina by comparing responses of a resting and a moving eye. The performance of simulated ganglion cells in a hyperacuity task is evaluated by ideal observer analysis. We ■nd that in the central retina eye-micromovements have no effect on the perfor- mance. Here optical blurring limits vernier acuity. In the retinal periph- ery however, eye-micromovements clearly improve performance. Based on ROC analysis, our predictions are quantitatively testable in electro- physiological and psychophysical experiments.
****
Mechanism of Neural Interference by Transcranial Magnetic Stimulation: Network or Single Neuron?
Yoichi Miyawaki, Masato Okada
This paper proposes neural mechanisms of transcranial magnetic stim- ulation (TMS). TMS can stimulate the brain non-invasively through a brief magnetic pulse delivered by a coil placed on the scalp, interfering with speci■c cortical functions with a high temporal resolution. Due to these advantages, TMS has been a popular experimental tool in various neuroscience ■elds. However, the neural mechanisms underlying TMS- induced interference are still unknown; a theoretical basis for TMS has not been developed. This paper provides computational evidence that in- hibitory interactions in a neural population, not an isolated single neuron, play a critical role in yielding the neural interference induced by TMS.
****
Efficient Multiscale Sampling from Products of Gaussian Mixtures
Alexander Ihler, Erik Sudderth, William Freeman, Alan Willsky

The problem of approximating the product of several Gaussian mixture distributio ns arises in a number of contexts, including the nonparametric belief propagatio n (NBP) inference algorithm and the training of prod- uct of experts models. Thi s paper develops two multiscale algorithms for sampling from a product of Gaussi an mixtures, and compares their performance to existing methods. The ■rst is a m ultiscale variant of pre- viously proposed Monte Carlo techniques, with comparab le theoretical guarantees but improved empirical convergence rates. The second m akes use of approximate kernel density evaluation methods to construct a fast ap proximate sampler, which is guaranteed to sample points to within a tunable para meter (cid:15) of their true probability. We compare both multi- scale samplers on a set of computational examples motivated by NBP, demonstrating signi■cant im provements over existing methods.

****

Markov Models for Automated ECG Interval Analysis

Nicholas Hughes, Lionel Tarassenko, Stephen J. Roberts

We examine the use of hidden Markov and hidden semi-Markov mod- els for automati cally segmenting an electrocardiogram waveform into its constituent waveform fea tures. An undecimated wavelet transform is used to generate an overcomplete repr esentation of the signal that is more appropriate for subsequent modelling. We s how that the state dura- tions implicit in a standard hidden Markov model are il l-suited to those of real ECG features, and we investigate the use of hidden sem i-Markov models for improved state duration modelling.

****

A Mixed-Signal VLSI for Real-Time Generation of Edge-Based Image Vectors

Masakazu Yagi, Hideo Yamasaki, Tadashi Shibata

A mixed-signal image filtering VLSI has been developed aiming at real-time gene ration of edge-based image vectors for robust image recognition. A four-stage a synchronous median detection architec- ture based on analog digital mixed-s ignal circuits has been intro- duced to determine the threshold value of edge detection, the key processing parameter in vector generation. As a result, a fully seamless pipeline processing from threshold detection to edge fea- ture map generation has been established. A prototype chip was designed in a 0.35-μm double-polysilicon three-metal-layer CMOS technolog y and the concept was verified by the fabricated chip. The chip generates a 64-dimension feature vector from a 64x64-pixel gray scale image every 80μs ec. This is about 104 times faster than the software computation, making a real -time image recognition system feasible.

****

Autonomous Helicopter Flight via Reinforcement Learning

H. Kim, Michael Jordan, Shankar Sastry, Andrew Ng

Autonomous helicopter ■ight represents a challenging control problem, with compl ex, noisy, dynamics. In this paper, we describe a successful application of rein forcement learning to autonomous helicopter ■ight. We ■rst ■t a stochastic, nonl inear model of the helicopter dynamics. We then use the model to learn to hover in place, and to ■y a number of maneuvers taken from an RC helicopter competitio n.

****

Insights from Machine Learning Applied to Human Visual Classification

Felix A. Wichmann, Arnulf Graf

We attempt to understand visual classi■cation in humans using both psy- chophysi cal and machine learning techniques. Frontal views of human faces were used for a gender classi■cation task. Human subjects classi- ■ed the faces and their gend er judgment, reaction time and con■dence rating were recorded. Several hyperplan e learning algorithms were used on the same classi■cation task using the Princip al Components of the texture and shape representation of the faces. The classi■c ation perfor- mance of the learning algorithms was estimated using the face data base with the true gender of the faces as labels, and also with the gender es- t imated by the subjects. We then correlated the human responses to the distance o f the stimuli to the separating hyperplane of the learning algo- rithms. Our res ults suggest that human classi■cation can be modeled by some hyperplane algorith

ms in the feature space we used. For classi■ca- tion, the brain needs more proce ssing for stimuli close to that hyperplane than for those further away.
****

Classification with Hybrid Generative/Discriminative Models
Rajat Raina, Yirong Shen, Andrew McCallum, Andrew Ng
Although discriminatively trained classi■ers are usually more accurate when labe led training data is abundant, previous work has shown that when training data i s limited, generative classi■ers can out-perform them. This paper describes a hy brid model in which a high-dimensional subset of the parameters are trained to m aximize generative likelihood, and another, small, subset of parameters are disc riminatively trained to maximize conditional likelihood. We give a sample comple xity bound showing that in order to ■t the discriminative parameters well, the n um- ber of training examples required depends only on the logarithm of the numbe r of feature occurrences and feature set size. Experimental results show that hy brid models can provide lower test error and can produce better accuracy/coverag e curves than either their purely generative or purely discriminative counterpar ts. We also discuss several advantages of hybrid models, and advocate further wo rk in this area.
****

Fast Feature Selection from Microarray Expression Data via Multiplicative Large Margin Algorithms
Claudio Gentile
New feature selection algorithms for linear threshold functions are de- scribed which combine backward elimination with an adaptive regular- ization method. Thi s makes them particularly suitable to the classi■ca- tion of microarray expressi on data, where the goal is to obtain accurate rules depending on few genes only. Our algorithms are fast and easy to implement, since they center on an incremen tal (large margin) algorithm which allows us to avoid linear, quadratic or highe r-order programming methods. We report on preliminary experiments with ■ve known DNA microarray datasets. These experiments suggest that multiplicative large ma rgin algorithms tend to outperform additive algorithms (such as SVM) on feature selection tasks.
****

Feature Selection in Clustering Problems
Volker Roth, Tilman Lange
A novel approach to combining clustering and feature selection is pre- sented. I t implements a wrapper strategy for feature selection, in the sense that the fea tures are directly selected by optimizing the discrimina- tive power of the used partitioning algorithm. On the technical side, we present an ef■cient optimizat ion algorithm with guaranteed local con- vergence property. The only free parame ter of this method is selected by a resampling-based stability analysis. Experim ents with real-world datasets demonstrate that our method is able to infer both meaningful partitions and meaningful subsets of features.
****

Optimal Manifold Representation of Data: An Information Theoretic Approach
Denis Chigirev, William Bialek
We introduce an information theoretic method for nonparametric, non- linear dime nsionality reduction, based on the in■nite cluster limit of rate distortion theo ry. By constraining the information available to manifold coordinates, a natural probabilistic map emerges that assigns original data to corresponding points on a lower dimensional manifold. With only the information-distortion trade off as a parameter, our method de- termines the shape of the manifold, its dimensional ity, the probabilistic map and the prior that provide optimal description of the data.
****

Invariant Pattern Recognition by Semi-Definite Programming Machines
Thore Graepel, Ralf Herbrich
invariances with respect to given pattern Knowledge about local transformations can greatly improve the accuracy of classi■cation. Previous approaches are eithe r based on regularisation or on the gen- eration of virtual (transformed) exampl

es. We develop a new frame- work for learning linear classi■ers under known tran sformations based on semide■nite programming. We present a new learning algorith m— the Semide■nite Programming Machine (SDPM)—which is able to ■nd a maximum mar gin hyperplane when the training examples are polynomial trajectories instead of single points. The solution is found to be sparse in dual variables and allows to identify those points on the trajectory with minimal real-valued output as vi rtual support vec- tors. Extensions to segments of trajectories, to more than on e trans- formation parameter, and to learning with kernels are discussed. In exp eriments we use a Taylor expansion to locally approximate rota- tional invarianc e in pixel images from USPS and ■nd improvements over known methods.

****

Impact of an Energy Normalization Transform on the Performance of the LF-ASD Bra in Computer Interface

Yu Zhou, Steven Mason, Gary Birch

This paper presents an energy normalization transform as a method to re duce system errors in the LF-ASD brain-computer interface. The energy normaliz ation transform has two major benefits to the system performance. First, it can increase class separation between the active and idle EEG data. Seco nd, it can desensitize the system to the signal amplitude variability. For four subjects in the study, the benefits resulted in the perfor mance improvement of the LF-ASD in the range from 7.7% to 18.9%, while for the fifth subject, who had the highest non-normalized accuracy of 90.5%, the performance did not change notably with normalization.

****

Subject-Independent Magnetoencephalographic Source Localization by a Multilayer Perceptron

Sung Jun, Barak Pearlmutter

We describe a system that localizes a single dipole to reasonable accu- racy fro m noisy magnetoencephalographic (MEG) measurements in real time. At its core is a multilayer perceptron (MLP) trained to map sen- sor signals and head position to dipole location. Including head position overcomes the previous need to retra in the MLP for each subject and ses- sion. The training dataset was generated by mapping randomly chosen dipoles and head positions through an analytic model an d adding noise from real MEG recordings. After training, a localization took 0.7 ms with an average error of 0.90 cm. A few iterations of a Levenberg-Marquardt routine using the MLP’s output as its initial guess took 15 ms and im- proved th e accuracy to 0.53 cm, only slightly above the statistical limits on accuracy im posed by the noise. We applied these methods to localize single dipole sources f rom MEG components isolated by blind source separation and compared the estimate d locations to those generated by standard manually-assisted commercial software .

****

Image Reconstruction by Linear Programming

Koji Tsuda, Gunnar Rätsch

A common way of image denoising is to project a noisy image to the sub- space of admissible images made for instance by PCA. However, a major drawback of this m ethod is that all pixels are updated by the projection, even when only a few pix els are corrupted by noise or occlusion. We pro- pose a new method to identify t he noisy pixels by (cid:1) 1-norm penalization and update the identi■ed pixels o nly. The identi■cation and updating of noisy pixels are formulated as one linear program which can be solved ef■ciently. Especially, one can apply the ν-trick t o directly specify the fraction of pixels to be reconstructed. Moreover, we exte nd the linear program to be able to exploit prior knowledge that occlusions ofte n ap- pear in contiguous blocks (e.g. sunglasses on faces). The basic idea is to penalize boundary points and interior points of the occluded area dif- ferently . We are able to show the ν-property also for this extended LP leading a method which is easy to use. Experimental results impressively demonstrate the power of our approach.

****

Extreme Components Analysis

Max Welling, Christopher Williams, Felix Agakov
Principal components analysis (PCA) is one of the most widely used techniques in machine learning and data mining. Minor components analysis (MCA) is less well known, but can also play an important role in the presence of constraints on the data distribution. In this paper we present a probabilistic model for “extreme components analysis” (XCA) which at the maximum likelihood solution extracts an optimal combina- tion of principal and minor components. For a given number of c ompo- nents, the log-likelihood of the XCA model is guaranteed to be larger or e qual than that of the probabilistic models for PCA and MCA. We de- scribe an ef■ cient algorithm to solve for the globally optimal solution. For log-convex spect ra we prove that the solution consists of principal components only, while for l og-concave spectra the solution consists of minor components. In general, the so lution admits a combination of both. In experiments we explore the properties of XCA on some synthetic and real-world datasets.
****

Bayesian Color Constancy with Non-Gaussian Models
Charles Rosenberg, Alok Ladsariya, Tom Minka
We present a Bayesian approach to color constancy which utilizes a non- Gaussian probabilistic model of the image formation process. The pa- rameters of this mo del are estimated directly from an uncalibrated image set and a small number of additional algorithmic parameters are chosen using cross validation. The algorit hm is empirically shown to exhibit RMS error lower than other color constancy al gorithms based on the Lambertian surface re■ectance model when estimating the il luminants of a set of test images. This is demonstrated via a direct performance comparison utilizing a publicly available set of real world test images and cod e base.
****

All learning is Local: Multi-agent Learning in Global Reward Games
Yu-han Chang, Tracey Ho, Leslie Kaelbling
In large multiagent games, partial observability, coordination, and credit assig nment persistently plague attempts to design good learning algo- rithms. We prov ide a simple and ef■cient algorithm that in part uses a linear system to model t he world from a single agent’s limited per- spective, and takes advantage of Kal man ■ltering to allow an agent to construct a good training signal and learn an effective policy.
****

Automatic Annotation of Everyday Movements
Deva Ramanan, David Forsyth
This paper describes a system that can annotate a video sequence with: a descrip tion of the appearance of each actor; when the actor is in view; and a represent ation of the actor’s activity while in view. The system does not require a ■xed background, and is automatic. The system works by (1) tracking people in 2D and then, using an annotated motion capture dataset, (2) synthesizing an annotated 3 D motion sequence matching the 2D tracks. The 3D motion capture data is manually annotated off-line using a class structure that describes everyday motions and allows mo- tion annotations to be composed — one may jump while running, for exa mple. Descriptions computed from video of real motions show that the method is a ccurate.
****

A Classification-based Cocktail-party Processor
Nicoleta Roman, Deliang Wang, Guy Brown
Guy J. Brown
****

Linear Dependent Dimensionality Reduction
Nathan Srebro, Tommi Jaakkola
We formulate linear dimensionality reduction as a semi-parametric esti- mation p roblem, enabling us to study its asymptotic behavior. We gen- eralize the proble m beyond additive Gaussian noise to (unknown) non- Gaussian additive noise, and to unbiased non-additive models.
****

Clustering with the Connectivity Kernel
Bernd Fischer, Volker Roth, Joachim Buhmann
Clustering aims at extracting hidden structure in dataset. While the prob- lem of ■nding compact clusters has been widely studied in the litera- ture, extracting arbitrarily formed elongated structures is considered a much harder problem. In this paper we present a novel clustering algo- rithm which tackles the problem by a two step procedure: ■rst the data are transformed in such a way that elongated structures become compact ones. In a second step, these new objects are clustered by optimizing a compactness-based criterion. The advantages of the method over related approaches are threefold: (i) robustness properties of compactness-based criteria naturally transfer to the problem of extracting elongated struc- tures, leading to a model which is highly robust against outlier objects; (ii) the transformed distances induce a Mercer kernel which allows us to formulate a polynomial approximation scheme to the generally NP- hard clustering problem; (iii) the new method does not contain free kernel parameters in contrast to methods like spectral clustering or mean-shift clustering.
****

Attractive People: Assembling Loose-Limbed Models using Non-parametric Belief Propagation
Leonid Sigal, Michael Isard, Benjamin Sigelman, Michael Black
The detection and pose estimation of people in images and video is made challenging by the variability of human appearance, the complexity of natural scenes, and the high dimensionality of articulated body mod- els. To cope with these problems we represent the 3D human body as a graphical model in which the relationships between the body parts are represented by conditional probability distributions. We formulate the pose estimation problem as one of probabilistic inference over a graphi- cal model where the random variables correspond to the individual limb parameters (position and orientation). Because the limbs are described by 6-dimensional vectors encoding pose in 3-space, discretization is im- practical and the random variables in our model must be continuous- valued. To approximate belief propagation in such a graph we exploit a recently introduced generalization of the particle ■lter. This framework facilitates the automatic initialization of the body-model from low level cues and is robust to occlusion of body parts and scene clutter.
****

Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering
Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, Marie Ouimet
Several unsupervised learning algorithms based on an eigendecompo- sition provide either an embedding or a clustering only for given train- ing points, with no straightforward extension for out-of-sample examples short of recomputing eigenvectors. This paper provides a uni■ed frame- work for extending Local Linear Embedding (LLE), Isomap, Laplacian Eigenmaps, Multi-Dimensional Scaling (for dimensionality reduction) as well as for Spectral Clustering. This framework is based on seeing these algorithms as learning eigenfunctions of a data-dependent kernel. Numerical experiments show that the generalizations performed have a level of error comparable to the variability of the embedding algorithms due to the choice of training data.
****

Learning Spectral Clustering
Francis Bach, Michael Jordan
Spectral clustering refers to a class of techniques which rely on the eigen- structure of a similarity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in dif- ferent clusters having low similarity. In this paper, we derive a new cost function for spectral clustering based on a measure of error between a given partition and a solution of the spectral relaxation of a minimum normalized cut problem. Minimizing this cost function with respect to the partition leads to a new spectral clustering algorithm. Minimizing with respect to the similarity matrix leads to an algor

ithm for learning the similarity matrix. We develop a tractable approximation of our cost function that is based on the power method of computing eigenvectors.

****

Maximum Likelihood Estimation of a Stochastic Integrate-and-Fire Neural Model
Liam Paninski, Eero Simoncelli, Jonathan Pillow

Recent work has examined the estimation of models of stimulus-driven neural activity in which some linear ■ltering process is followed by a nonlinear, probabilistic spiking stage. We analyze the estimation of one such model for which this nonlinear step is implemented by a noisy, leaky, integrate-and-■re mechanism with a spike-dependent after- current. This model is a biophysically plausible alternative to models with Poisson (memory-less) spiking, and has been shown to effectively reproduce various spiking statistics of neurons in vivo. However, the problem of estimating the model from extracellular spike train data has not been examined in depth. We formulate the problem in terms of max- imum likelihood estimation, and show that the computational problem of maximizing the likelihood is tractable. Our main contribution is an algorithm and a proof that this algorithm is guaranteed to ■nd the global optimum with reasonable speed. We demonstrate the effectiveness of our estimator with numerical simulations.

****

Prediction on Spike Data Using Kernel Algorithms
Jan Eichhorn, Andreas Tolias, Alexander Zien, Malte Kuss, Jason Weston, Nikos Logothetis, Bernhard Schölkopf, Carl Rasmussen

We report and compare the performance of different learning algorithms based on data from cortical recordings. The task is to predict the orienta- tion of visual stimuli from the activity of a population of simultaneously recorded neurons. We compare several ways of improving the coding of the input (i.e., the spike data) as well as of the output (i.e., the orienta- tion), and report the results obtained using different kernel algorithms.

****

An MDP-Based Approach to Online Mechanism Design
David C. Parkes, Satinder Singh

Online mechanism design (MD) considers the problem of provid- ing incentives to implement desired system-wide outcomes in sys- tems with self-interested agents that arrive and depart dynami- cally. Agents can choose to misrepresent their arrival and depar- ture times, in addition to information about their value for di(cid:11)erent outcomes. We consider the problem of maximizing the total long- term value of the system despite the self-interest of agents. The online MD problem induces a Markov Decision Process (MDP), which when solved can be used to implement optimal policies in a truth-revealing Bayesian-Nash equilibrium.

****

Probabilistic Inference of Speech Signals from Phaseless Spectrograms
Kannan Achan, Sam Roweis, Brendan J. Frey

Many techniques for complex speech processing such as denoising and deconvolution, time/frequency warping, multiple speaker separation, and multiple microphone analysis operate on sequences of short-time power spectra (spectrograms), a representation which is often well-suited to these tasks. However, a signi■cant problem with algorithms that manipu- late spectrograms is that the output spectrogram does not include a phase component, which is needed to create a time-domain signal that has good perceptual quality. Here we describe a generative model of time-domain speech signals and their spectrograms, and show how an ef■cient optimizer can be used to ■nd the maximum a posteriori speech signal, given the spectrogram. In contrast to techniques that alternate between esti- mating the phase and a spectrally-consistent signal, our technique di- rectly infers the speech signal, thus jointly optimizing the phase and a spectrally-consistent signal. We compare our technique with a standard method using signal-to-noise ratios, but we also provide audio ■les on the web for the purpose of demonstrating the improvement in perceptual quality that our technique offers.

****

Denoising and Untangling Graphs Using Degree Priors
Quaid Morris, Brendan J. Frey

This paper addresses the problem of untangling hidden graphs from a set of noisy detections of undirected edges. We present a model of the generation of the observed graph that includes degree-based structure priors on the hidden graphs. Exact inference in the model is intractable; we present an e–cient approximate inference algo- rithm to compute edge appearance posteriors. We evaluate our model and algorithm on a biological graph inference problem.

****

Learning a Distance Metric from Relative Comparisons

Matthew Schultz, Thorsten Joachims

This paper presents a method for learning a distance metric from rel- ative comparison such as “A is closer to B than A is to C”. Taking a Support Vector Machine (SVM) approach, we develop an algorithm that provides a ■exible way of describing qualitative training data as a set of constraints. We show that such constraints lead to a convex quadratic programming problem that can be solved by adapting standard meth- ods for SVM training. We empirically evaluate the performance and the modelling ■exibility of the algorithm on a collection of text documents.

****

Locality Preserving Projections

Xiaofei He, Partha Niyogi

Many problems in information processing involve some form of dimen- sionality reduction. In this paper, we introduce Locality Preserving Pro- jections (LPP). These are linear projective maps that arise by solving a variational problem that optimally preserves the neighborhood structure of the data set. LPP should be seen as an alternative to Principal Com- ponent Analysis (PCA) – a classical linear technique that projects the data along the directions of maximal variance. When the high dimen- sional data lies on a low dimensional manifold embedded in the ambient space, the Locality Preserving Projections are obtained by ■nding the optimal linear approximations to the eigenfunctions of the Laplace Bel- trami operator on the manifold. As a result, LPP shares many of the data representation properties of nonlinear techniques such as Laplacian Eigenmaps or Locally Linear Embedding. Yet LPP is linear and more crucially is de■ned everywhere in ambient space rather than just on the training data points. This is borne out by illustrative examples on some high dimensional data sets.

****

Unsupervised Color Decomposition Of Histologically Stained Tissue Samples

Andrew Rabinovich, Sameer Agarwal, Casey Laris, Jeffrey Price, Serge Belongie

Accurate spectral decomposition is essential for the analysis and diagnosis of histologically stained tissue sections. In this paper we present the ■rst automated system for performing this decompo- sition. We compare the performance of our system with ground truth data and report favorable results.

****

Minimising Contrastive Divergence in Noisy, Mixed-mode VLSI Neurons

Hsin Chen, Patrice Fleury, Alan Murray

This paper presents VLSI circuits with continuous-valued proba- bilistic behaviour realized by injecting noise into each computing unit(neuron). Interconnecting the noisy neurons forms a Contin- uous Restricted Boltzmann Machine (CRBM), which has shown promising performance in modelling and classifying noisy biomed- ical data. The Minimising-Contrastive-Divergence learning algo- rithm for CRBM is also implemented in mixed-mode VLSI, to adapt the noisy neurons’ parameters on-chip.

****

Eye Movements for Reward Maximization

Nathan Sprague, Dana Ballard

University of Rochester Rochester, NY 14627 dana@cs.rochester.edu

****

Estimating Internal Variables and Paramters of a Learning Agent by a Particle Filter

Kazuyuki Samejima, Kenji Doya, Yasumasa Ueda, Minoru Kimura

When we model a higher order functions, such as learning and memory, we face a dif■culty of comparing neural activities with hidden variables that depend on the

history of sensory and motor signals and the dynam- ics of the network. Here, we propose novel method for estimating hidden variables of a learning agent, such as connection weights from sequences of observable variables. Bayesian estimation is a method to estimate the posterior probability of hidden variables from observable data sequence using a dynamic model of hidden and observable variables. In this pa- per, we apply particle ■lter for estimating internal parameters and meta- parameters of a reinforcement learning model. We veri■ed the effective- ness of the method using both arti■cial data and real animal behavioral data.

****

ICA-based Clustering of Genes from Microarray Expression Data

Su-in Lee, Serafim Batzoglou

We propose an unsupervised methodology using independent component analysis (ICA) to cluster genes from DNA microarray data. Based on an ICA mixture model of genomic expression patterns, linear and nonlinear ICA finds components that are specific to certain biological processes. Genes that exhibit significant up-regulation or down-regulation within each component are grouped into clusters. We test the statistical significance of enrichment of gene annotations within each cluster. ICA-based clustering outperformed other leading methods in constructing functionally coherent clusters on various datasets. This result supports our model of genomic expression data as composite effect of independent biological processes. Comparison of clustering performance among various including a kernel-based nonlinear ICA algorithm shows that nonlinear ICA performed the best for small datasets and natural-gradient maximization-likelihood worked well for all the datasets.

****

On the Concentration of Expectation and Approximate Inference in Layered Networks

XuanLong Nguyen, Michael Jordan

We present an analysis of concentration-of-expectation phenomena in layered Bayesian networks that use generalized linear models as the local conditional probabilities. This framework encompasses a wide variety of probability distributions, including both discrete and continuous random variables. We utilize ideas from large deviation analysis and the delta method to devise and evaluate a class of approximate inference algo- rithms for layered Bayesian networks that have superior asymptotic error bounds and very fast computation time.

****

Identifying Structure across Pre-partitioned Data

Zvika Marx, Ido Dagan, Eli Shamir

We propose an information-theoretic clustering approach that incorporates a pre-known partition of the data, aiming to identify common clusters that cut across the given partition. In the standard clustering setting the formation of clusters is guided by a single source of feature information. The newly utilized pre-partition factor introduces an additional bias that counterbalances the impact of the features whenever they become correlated with this known partition. The resulting algorithmic framework was applied successfully to synthetic data, as well as to identifying text-based cross-religion correspondences.

****

Eigenvoice Speaker Adaptation via Composite Kernel Principal Component Analysis

James Kwok, Brian Mak, Simon Ho

Eigenvoice speaker adaptation has been shown to be effective when only a small amount of adaptation data is available. At the heart of the method is principal component analysis (PCA) employed to ■nd the most im- portant eigenvoices. In this paper, we postulate that nonlinear PCA, in particular kernel PCA, may be even more effective. One major challenge is to map the feature-space eigenvoices back to the observation space so that the state observation likelihoods can be computed during the estima- tion of eigenvoice weights and subsequent decoding. Our solution is to compute kernel PCA using composite kernels, and we will call our new method kernel eigenvoice speaker adaptation. On the TIDIGITS corpus, we found that compared with a speaker-independent model, our kernel eigenvoice adaptatio

n method can reduce the word error rate by 28–33% while the standard eigenvoice approach can only match the performance of the speaker-independent model.
****

A Summating, Exponentially-Decaying CMOS Synapse for Spiking Neural Systems
Z. Shi, Timothy Horiuchi
Synapses are a critical element of biologically-realistic, spike-based neu- ral computation, serving the role of communication, computation, and modi■cation. Many different circuit implementations of synapse func- tion exist with different computational goals in mind. In this paper we describe a new CMOS synapse design that separately controls quiescent leak current, synaptic gain, and time-constant of decay. This circuit im- plements part of a commonly-used kinetic model of synaptic conduc- tance. We show a theoretical analysis and experimental data for proto- types fabricated in a commercially-available 1.5µm CMOS process.
****

Extending Q-Learning to General Adaptive Multi-Agent Systems
Gerald Tesauro
Recent multi-agent extensions of Q-Learning require knowledge of other agents' payoffs and Q-functions, and assume game-theoretic play at all times by all other agents. This paper proposes a fundamentally different approach, dubbed "Hyper-Q" Learning, in which values of mixed strategies rather than base actions are learned, and in which other agents' strategies are estimated from observed actions via Bayesian in- ference. Hyper-Q may be effective against many different types of adap- tive agents, even if they are persistently dynamic. Against certain broad categories of adaptation, it is argued that Hyper-Q may converge to ex- act optimal time-varying policies. In tests using Rock-Paper-Scissors, Hyper-Q learns to significantly exploit an Infinitesimal Gradient Ascent (IGA) player, as well as a Policy Hill Climber (PHC) player. Preliminary analysis of Hyper-Q against itself is also presented.
****

No Unbiased Estimator of the Variance of K-Fold Cross-Validation
Yoshua Bengio, Yves Grandvalet
Most machine learning researchers perform quantitative experiments to estimate generalization error and compare algorithm performances. In order to draw statistically convincing conclusions, it is important to esti- mate the uncertainty of such estimates. This paper studies the estimation of uncertainty around the K-fold cross-validation estimator. The main theorem shows that there exists no universal unbiased estimator of the variance of K-fold cross-validation. An analysis based on the eigende- composition of the covariance matrix of errors helps to better understand the nature of the problem and shows that naive estimators may grossly underestimate variance, as con£rmed by numerical experiments.
****

GPPS: A Gaussian Process Positioning System for Cellular Networks
Anton Schwaighofer, Marian Grigoras, Volker Tresp, Clemens Hoffmann
In this article, we present a novel approach to solving the localization problem in cellular networks. The goal is to estimate a mobile user's position, based on measurements of the signal strengths received from network base stations. Our solution works by building Gaussian process models for the distribution of signal strengths, as obtained in a series of calibration measurements. In the localization stage, the user's posi- tion can be estimated by maximizing the likelihood of received signal strengths with respect to the position. We investigate the accuracy of the proposed approach on data obtained within a large indoor cellular network.
****

Reconstructing MEG Sources with Unknown Correlations
Maneesh Sahani, Srikantan Nagarajan
Existing source location and recovery algorithms used in magnetoen- cephalographic imaging generally assume that the source activity at dif- ferent brain locations is independent or that the correlation structure is known. However, electrophysiological recordings of local ■eld poten- tials show strong correlations in aggregate activity over signi■cant dis- tances. Indeed, it seems very likely that

stimulus-evoked activity would follow strongly correlated time-courses in diffe rent brain areas. Here, we present, and validate through simulations, a new appr oach to source reconstruction in which the correlation between sources is modell ed and estimated explicitly by variational Bayesian methods, facilitating accu- rate recovery of source locations and the time-courses of their activation.
****

Increase Information Transfer Rates in BCI by CSP Extension to Multi-class
Guido Dornhege, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller
Brain-Computer Interfaces (BCI) are an interesting emerging technology that is d riven by the motivation to develop an effective communication in- terface transl ating human intentions into a control signal for devices like computers or neuro prostheses. If this can be done bypassing the usual hu- man output pathways like peripheral nerves and muscles it can ultimately become a valuable tool for para lyzed patients. Most activity in BCI re- search is devoted to ■nding suitable fe atures and algorithms to increase information transfer rates (ITRs). The present paper studies the implica- tions of using more classes, e.g., left vs. right ha nd vs. foot, for operating a BCI. We contribute by (1) a theoretical study showi ng under some mild assumptions that it is practically not useful to employ more than three or four classes, (2) two extensions of the common spatial pattern (CS P) algorithm, one interestingly based on simultaneous diagonalization, and (3) c ontrolled EEG experiments that underline our theoretical ■ndings and show excell ent improved ITRs.
****

Dopamine Modulation in a Basal Ganglio-Cortical Network of Working Memory
Aaron Gruber, Peter Dayan, Boris Gutkin, Sara Solla
Dopamine exerts two classes of effect on the sustained neural activity in prefro ntal cortex that underlies working memory. Direct release in the cortex increase s the contrast of prefrontal neurons, enhancing the ro- bustness of storage. Rel ease of dopamine in the striatum is associated with salient stimuli and makes me dium spiny neurons bistable; this mod- ulation of the output of spiny neurons af fects prefrontal cortex so as to indirectly gate access to working memory and ad ditionally damp sensi- tivity to noise. Existing models have treated dopamine in one or other structure, or have addressed basal ganglia gating of working memor y ex- clusive of dopamine effects. In this paper we combine these mechanisms and explore their joint effect. We model a memory-guided saccade task to illustrate how dopamine’s actions lead to working memory that is se- lective for salient i nput and has increased robustness to distraction.
****

Approximate Planning in POMDPs with Macro-Actions
Georgios Theocharous, Leslie Kaelbling
Leslie Pack Kaelbling
****

ARA*: Anytime A* with Provable Bounds on Sub-Optimality
Maxim Likhachev, Geoffrey J. Gordon, Sebastian Thrun
In real world planning problems, time for deliberation is often limited. Anytime planners are well suited for these problems: they ■nd a feasi- ble solution qui ckly and then continually work on improving it until time runs out. In this pape r we propose an anytime heuristic search, ARA, which tunes its performance bound based on available search time. It starts by ■nding a suboptimal solution quick ly using a loose bound, then tightens the bound progressively as time allows. Gi ven enough time it ■nds a provably optimal solution. While improving its bound, ARA reuses previous search efforts and, as a result, is signi■cantly more ef■- c ient than other anytime search methods. In addition to our theoretical analysis, we demonstrate the practical utility of ARA* with experiments on a simulated ro bot kinematic arm and a dynamic path planning prob- lem for an outdoor rover.
****

Salient Boundary Detection using Ratio Contour
Song Wang, Toshiro Kubota, Jeffrey Siskind
This paper presents a novel graph-theoretic approach, named ratio con- tour, to extract perceptually salient boundaries from a set of noisy bound- ary fragments

detected in real images. The boundary saliency is de■ned using the Gestalt laws of closure, proximity, and continuity. This pa- per ■rst constructs an undirected graph with two different sets of edges: solid edges and dashed edges. The weights of solid and dashed edges measure the local saliency in and between boundary fragments, respec- tively. Then the most salient boundary is detected by searching for an optimal cycle in this graph with minimum average weight. The proposed approach guarantees the global optimality without introducing any biases related to region area or boundary length. We collect a variety of images for testing the proposed approach with encouraging results.

****

Semi-Definite Programming by Perceptron Learning

Thore Graepel, Ralf Herbrich, Andriy Kharechko, John Shawe-taylor

We present a modi■ed version of the perceptron learning algorithm (PLA) which solves semide■nite programs (SDPs) in polynomial time. The algorithm is based on the following three observations: (i) Semide■nite programs are linear programs with in■nitely many (linear) constraints; (ii) every linear program can be solved by a sequence of constraint satisfaction problems with linear constraints; (iii) in general, the perceptron learning algorithm solves a constraint satisfaction problem with linear constraints in ■nitely many updates. Combining the PLA with a probabilistic rescaling algorithm (which, on average, increases the size of the feasable region) results in a prob- abilistic algorithm for solving SDPs that runs in polynomial time. We present preliminary results which demonstrate that the algo- rithm works, but is not competitive with state-of-the-art interior point methods.

****

An Iterative Improvement Procedure for Hierarchical Clustering

David Kauchak, Sanjoy Dasgupta

We describe a procedure which ■nds a hierarchical clustering by hill- climbing. The cost function we use is a hierarchical extension of the k-means cost; our local moves are tree restructurings and node reorder- ings. We show these can be accomplished ef■ciently, by exploiting spe- cial properties of squared Euclidean distances and by using techniques from scheduling algorithms.

****

Kernels for Structured Natural Language Data

Jun Suzuki, Yutaka Sasaki, Eisaku Maeda

This paper devises a novel kernel function for structured natural language data. In the ■eld of Natural Language Processing, feature extraction consists of the following two steps: (1) syntactically and semantically analyzing raw data, i.e., character strings, then representing the results as discrete structures, such as parse trees and dependency graphs with part-of-speech tags; (2) creating (possibly high-dimensional) numerical feature vectors from the discrete structures. The new kernels, called Hier- archical Directed Acyclic Graph (HDAG) kernels, directly accept DAGs whose nodes can contain DAGs. HDAG data structures are needed to fully re■ect the syntactic and semantic structures that natural language data inherently have. In this paper, we de■ne the kernel function and show how it permits ef■cient calculation. Experiments demonstrate that the proposed kernels are superior to existing kernel functions, e.g., se- quence kernels, tree kernels, and bag-of-words kernels.

****

Near-Minimax Optimal Classification with Dyadic Classification Trees

Clayton Scott, Robert Nowak

This paper reports on a family of computationally practical classi■ers that converge to the Bayes error at near-minimax optimal rates for a va- riety of distributions. The classi■ers are based on dyadic classi■cation trees (DCTs), which involve adaptively pruned partitions of the feature space. A key aspect of DCTs is their spatial adaptivity, which enables lo- cal (rather than global) ■tting of the decision boundary. Our risk analysis involves a spatial decomposition of the usual concentration inequalities, leading to a spatially adaptive, data-dependent pruning criterion. For any distribution on (X, Y) whose Bayes decision boundary behaves locally like a Lipschitz smooth function, we show that the DCT error

converges to the Bayes error at a rate within a logarithmic factor of the minima x optimal rate. We also study DCTs equipped with polynomial classi■ca- tion rule s at each leaf, and show that as the smoothness of the boundary increases their errors converge to the Bayes error at a rate approaching n—1/2, the parametric r ate. We are not aware of any other practical classi- ■ers that provide similar r ate of convergence guarantees. Fast algorithms for tree pruning are discussed.
****

Bounded Invariance and the Formation of Place Fields
Reto Wyss, Paul Verschure
One current explanation of the view independent representation of space by the p lace-cells of the hippocampus is that they arise out of the summation of view de pendent Gaussians. This proposal as- sumes that visual representations show boun ded invariance. Here we investigate whether a recently proposed visual encoding scheme called the temporal population code can provide such representa- tions. O ur analysis is based on the behavior of a simulated robot in a virtual environme nt containing speci(cid:12)c visual cues. Our re- sults show that the temporal p opulation code provides a represen- tational substrate that can naturally accoun t for the formation of place (cid:12)elds.
****

Learning Curves for Stochastic Gradient Descent in Linear Feedforward Networks
Justin Werfel, Xiaohui Xie, H. Seung
Dept. of Brain & Cog. Sci.
****

Learning Bounds for a Generalized Family of Bayesian Posterior Distributions
Tong Zhang
In this paper we obtain convergence bounds for the concentration of Bayesian pos terior distributions (around the true distribution) using a novel method that si mpli■es and enhances previous results. Based on the analysis, we also introduce a generalized family of Bayesian posteriors, and show that the convergence behav ior of these generalized posteriors is completely determined by the local prior structure around the true distri- bution. This important and surprising robustne ss property does not hold for the standard Bayesian posterior in that it may not concentrate when there exist “bad” prior structures even at places far away fro m the true distribution.
****

A Functional Architecture for Motion Pattern Processing in MSTd
Scott Beardsley, Lucia Vaina
Lucia M. Vaina
****

Online Learning of Non-stationary Sequences
Claire Monteleoni, Tommi Jaakkola
We consider an online learning scenario in which the learner can make prediction s on the basis of a ■xed set of experts. We derive upper and lower relative loss bounds for a class of universal learning algorithms in- volving a switching dyn amics over the choice of the experts. On the basis of the performance bounds we provide the optimal a priori discretiza- tion for learning the parameter that go verns the switching dynamics. We demonstrate the new algorithm in the context of wireless networks.
****