An Analysis of Convex Relaxations for MAP Estimation

Pawan Mudigonda, Vladimir Kolmogorov, Philip Torr

The problem of obtaining the maximum a posteriori estimate of a general discrete random field (i.e. a random field defined using a finite and discrete set of labels) is known to be NP-hard. However, due to its central importance in many applications, several approximate algorithms have been proposed in the literature. In this paper, we present an analysis of three such algorithms based on convex relaxations: (i) LP-S: the linear programming (LP) relaxation proposed by Schlesinger [20] for a special case and independently in [4, 12, 23] for the general case; (ii) QP-RL: the quadratic programming (QP) relaxation by Ravikumar and Lafferty [18]; and (iii) SOCP-MS: the second order cone programming (SOCP) relaxation first proposed by Muramatsu and Suzuki [16] for two label problems and later extended in [14] for a general label set. We show that the SOCP-MS and the QP-RL relaxations are equivalent. Furthermore, we prove that despite the flexibility in the form of the constraints/objective function offered by QP and SOCP, the LP-S relaxation strictly dominates (i.e. provides a better approximation than) QP-RL and SOCP-MS. We generalize these results by defining a large class of SOCP (and equivalent QP) relaxations which is dominated by the LP-S relaxation. Based on these results we propose some novel SOCP relaxations which strictly dominate the previous approaches.

****

Random Features for Large-Scale Kernel Machines

Ali Rahimi, Benjamin Recht

To accelerate the training of kernel machines, we propose to map the input data to a randomized low-dimensional feature space and then apply existing fast linear methods. The features are designed so that the inner products of the transformed data are approximately equal to those in the feature space of a user specified shift- invariant kernel. We explore two sets of random features, provide convergence bounds on their ability to approximate various radial basis kernels, and show that in large-scale classification and regression tasks linear machine learning al- gorithms applied to these features outperform state-of-the-art large-scale kernel machines.

****

Compressed Regression

Shuheng Zhou, Larry Wasserman, John Lafferty

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-authors prior to requesting a name change in the electronic proceedings.

****

Simulated Annealing: Rigorous finite-time guarantees for optimization on continuous domains

Andrea Lecchini-visintini, John Lygeros, Jan Maciejowski

Simulated annealing is a popular method for approaching the solution of a global optimization problem. Existing results on its performance apply to discrete combinatorial optimization where the optimization variables can assume only a finite set of possible values. We introduce a new general formulation of simulated an- nealing which allows one to guarantee finite-time performance in the optimization of functions of continuous variables. The results hold universally for any optimization problem on a bounded domain and establish a connection between simulated annealing and up-to-date theory of convergence of Markov chain Monte Carlo methods on continuous domains. This work is inspired by the concept of finite-time learning with known accuracy and confidence developed in statistical learning theory.

****

Predictive Matrix-Variate t Models

Shenghuo Zhu, Kai Yu, Yihong Gong

It is becoming increasingly important to learn from a partially-observed random matrix and predict its missing elements. We assume that the entire matrix is a single sample drawn from a matrix-variate t distribution and suggest a matrix-var iate t model (MVTM) to predict those missing elements. We show that MVTM generalizes a range of known probabilistic models, and automatically performs model selection to encourage sparse predictive models. Due to the non-conjugacy of its prior, it is difficult to make predictions by computing the mode or mean of the posterior distribution. We suggest an optimization method that sequentially minimizes a convex upper-bound of the log-likelihood, which is very efficient and scalable. The experiments on a toy data and EachMovie dataset show a good predictive accuracy of the model.

****

Loop Series and Bethe Variational Bounds in Attractive Graphical Models

Alan Willsky, Erik Sudderth, Martin J. Wainwright

Variational methods are frequently used to approximate or bound the partition or likelihood function of a Markov random field. Methods based on mean field theory are guaranteed to provide lower bounds, whereas certain types of convex relaxations provide upper bounds. In general, loopy belief propagation (BP) provides (often accurate) approximations, but not bounds. We prove that for a class of attractive binary models, the value specified by any fixed point of loopy BP always provides a lower bound on the true likelihood. Empirically, this bound is much better than the naive mean field bound, and requires no further work than running BP. We establish these lower bounds using a loop series expansion due to Chertkov and Chernyak, which we show can be derived as a consequence of the tree reparameterization characterization of BP fixed points.

****

Stable Dual Dynamic Programming

Tao Wang, Michael Bowling, Dale Schuurmans, Daniel Lizotte

Recently, we have introduced a novel approach to dynamic programming and re- inforcement learning that is based on maintaining explicit representations of stationary distributions instead of value functions. In this paper, we investigate the convergence properties of these dual algorithms both theoretically and empirically, and show how they can be scaled up by incorporating function approximation.

****

FilterBoost: Regression and Classification on Large Datasets

Joseph K. Bradley, Robert E. Schapire

We study boosting in the filtering setting, where the booster draws examples from an oracle instead of using a fixed training set and so may train efficiently on very large datasets. Our algorithm, which is based on a logistic regression technique proposed by Collins, Schapire, & Singer, requires fewer assumptions to achieve bounds equivalent to or better than previous work. Moreover, we give the first proof that the algorithm of Collins et al. is a strong PAC learner, albeit within the filtering setting. Our proofs demonstrate the algorithm’s strong theoretical proper- ties for both classification and conditional probability estimation, and we validate these results through extensive experiments. Empirically, our algorithm proves more robust to noise and overfitting than batch boosters in conditional probability estimation and proves competitive in classification.

****

Unsupervised Feature Selection for Accurate Recommendation of High-Dimensional Image Data

Sabri Boutemedjet, Djemel Ziou, Nizar Bouguila

Content-based image suggestion (CBIS) targets the recommendation of products based on user preferences on the visual content of images. In this paper, we mo- tivate both feature selection and model order identification as two key issues for a successful CBIS. We propose a generative model in which the visual features and users are clustered into separate classes. We identify the number of both user and image classes with the simultaneous selection of relevant visual features us- ing the message length approach. The goal is to ensure an accurate prediction of ratings for multidimensional non-Gaussian and continuous image descriptors. Experiments on a collected data have demonstrated the merits of our approach.

****

Efficient Principled Learning of Thin Junction Trees

Anton Chechetka, Carlos Guestrin

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-authors prior to requesting a name change in the electronic proceedings.

****

Regret Minimization in Games with Incomplete Information

Martin Zinkevich, Michael Johanson, Michael Bowling, Carmelo Piccione

Extensive games are a powerful model of multiagent decision-making scenarios with incomplete information. Finding a Nash equilibrium for very large instances of these games has received a great deal of recent attention. In this paper, we describe a new technique for solving large games based on regret minimization. In particular, we introduce the notion of counterfactual regret, which exploits the degree of incomplete information in an extensive game. We show how minimizing counterfactual regret minimizes overall regret, and therefore in self-play can be used to compute a Nash equilibrium. We demonstrate this technique in the domain of poker, showing we can solve abstractions of limit Texas Hold’em with as many as 1012 states, two orders of magnitude larger than previous methods.

****

A Bayesian Model of Conditioned Perception

Alan A. Stocker, Eero Simoncelli

We propose an extended probabilistic model for human perception. We argue that in many circumstances, human observers simultaneously evaluate sensory evidence under different hypotheses regarding the underlying physical process that might have generated the sensory information. Within this context, inference can be optimal if the observer weighs each hypothesis according to the correct belief in that hypothesis. But if the observer commits to a particular hypothesis, the belief in that hypothesis is converted into subjective certainty, and subsequent perceptual behavior is suboptimal, conditioned only on the chosen hypothesis. We demonstrate that this framework can explain psychophysical data of a recently reported decision-estimation experiment. The model well accounts for the data, predicting the same estimation bias as a consequence of the preceding decision step. The power of the framework is that it has no free parameters except the degree of the observer's uncertainty about its internal sensory representation. All other parameters are defined by the particular experiment which allows us to make quantitative predictions of human perception to two modifications of the original experiment.

****

Scan Strategies for Meteorological Radars

Victoria Manfredi, Jim Kurose

We address the problem of adaptive sensor control in dynamic resource-constrained sensor networks. We focus on a meteorological sensing network comprising radars that can perform sector scanning rather than always scanning 360 degrees. We compare three sector scanning strategies. The sit-and-spin strategy always scans 360 degrees. The limited lookahead strategy additionally uses the expected environmental state K decision epochs in the future, as predicted from Kalman filters, in its decision-making. The full lookahead strategy uses all expected future states by casting the problem as a Markov decision process and using reinforcement learning to estimate the optimal scan strategy. We show that the main benefits of using a lookahead strategy are when there are multiple meteorological phenomena in the environment, and when the maximum radius of any phenomenon is sufficiently smaller than the radius of the radars. We also show that there is a trade-off between the average quality with which a phenomenon is scanned and the number of decision epochs before which a phenomenon is rescanned.

****

The Tradeoffs of Large Scale Learning

Léon Bottou, Olivier Bousquet

This contribution develops a theoretical framework that takes into account the effect of approximate optimization on learning algorithms. The analysis shows distinct tradeoffs for the case of small-scale and large-scale learning problems. S mall-scale learning problems are subject to the usual approximation--estimation tradeoff. Large-scale learning problems are subject to a qualitatively different tradeoff involving the computational complexity of the underlying optimization algorithms in non-trivial ways.

****

Inferring Elapsed Time from Stochastic Neural Processes

Misha Ahrens, Maneesh Sahani

Many perceptual processes and neural computations, such as speech recognition, motor control and learning, depend on the ability to measure and mark the passage of time. However, the processes that make such temporal judgements possible are unknown. A number of different hypothetical mechanisms have been advanced, all of which depend on the known, temporally predictable evolution of a neural or psychological state, possibly through oscillations or the gradual decay of a memory trace. Alternatively, judgements of elapsed time might be based on observations of temporally structured, but stochastic processes. Such processes need not be specific to the sense of time; typical neural and sensory processes contain at least some statistical structure across a range of time scales. Here, we investigate the statistical properties of an estimator of elapsed time which is based on a simple family of stochastic process.

****

A learning framework for nearest neighbor search

Lawrence Cayton, Sanjoy Dasgupta

Can we leverage learning techniques to build a fast nearest-neighbor (NN) retrieval data structure? We present a general learning framework for the NN problem in which sample queries are used to learn the parameters of a data structure that minimize the retrieval time and/or the miss rate. We explore the potential of this novel framework through two popular NN data structures: KD-trees and the rectilinear structures employed by locality sensitive hashing. We derive a generalization theory for these data structure classes and present simple learning algorithms for both. Experimental results reveal that learning often improves on the already strong performance of these data structures.

****

Reinforcement Learning in Continuous Action Spaces through Sequential Monte Carlo Methods

Alessandro Lazaric, Marcello Restelli, Andrea Bonarini

Learning in real-world domains often requires to deal with continuous state and action spaces. Although many solutions have been proposed to apply Reinforce- ment Learning algorithms to continuous state problems, the same techniques can be hardly extended to continuous action spaces, where, besides the computation of a good approximation of the value function, a fast method for the identification of the highest-valued action is needed. In this paper, we propose a novel actor-critic approach in which the policy of the actor is estimated through sequential Monte Carlo methods. The importance sampling step is performed on the basis of the values learned by the critic, while the resampling step modifies the actor’s policy. The proposed approach has been empirically compared to other learning algo- rithms into several domains; in this paper, we report results obtained in a control problem consisting of steering a boat across a river.

****

Ensemble Clustering using Semidefinite Programming

Vikas Singh, Lopamudra Mukherjee, Jiming Peng, Jinhui Xu

We consider the ensemble clustering problem where the task is to ‘aggregate’ multiple clustering solutions into a single consolidated clustering that maximizes the shared information among given clustering solutions. We obtain several new results for this problem. First, we note that the notion of agreement under such circumstances can be better captured using an agreement measure based on a 2D string encoding rather than voting strategy based methods proposed in literature. Using this generalization, we first derive a nonlinear optimization model to max- imize the new agreement measure. We then show that our optimization problem can be transformed into a strict 0-1 Semidefinite Program (SDP) via novel con- vexification techniques which can subsequently be relaxed to a polynomial time solvabl e SDP. Our experiments indicate improvements not only in terms of the proposed agreement measure but also the existing agreement measures based on voting strategies. We discuss evaluations on clustering and image segmentation databases.

****

Theoretical Analysis of Heuristic Search Methods for Online POMDPs
Stephane Ross, Joelle Pineau, Brahim Chaib-draa
Planning in partially observable environments remains a challenging problem, despite significant recent advances in offline approximation techniques. A few online methods have also been proposed recently, and proven to be remarkably scalable, but without the theoretical guarantees of their offline counterparts. Thus it seems natural to try to unify offline and online techniques, preserving the theoretical properties of the former, and exploiting the scalability of the latter. In this paper, we provide theoretical guarantees on an anytime algorithm for POMDPs which aims to reduce the error made by approximate offline value iteration algorithms through the use of an efficient online searching procedure. The algorithm uses search heuristics based on an error analysis of lookahead search, to guide the online search towards reachable beliefs with the most potential to reduce error. We provide a general theorem showing that these search heuristics are admissible, and lead to complete and epsilon-optimal algorithms. This is, to the best of our knowledge, the strongest theoretical result available for online POMDP solution methods. We also provide empirical evidence showing that our approach is also practical, and can find (provably) near-optimal solutions in reasonable time.

****

A Constraint Generation Approach to Learning Stable Linear Dynamical Systems
Byron Boots, Geoffrey J. Gordon, Sajid Siddiqi
Stability is a desirable characteristic for linear dynamical systems, but it is often ignored by algorithms that learn these systems from data. We propose a novel method for learning stable linear dynamical systems: we formulate an approxima- tion of the problem as a convex program, start with a solution to a relaxed version of the program, and incrementally add constraints to improve stability. Rather than continuing to generate constraints until we reach a feasible solution, we test stability at each step; because the convex program is only an approximation of the desired problem, this early stopping rule can yield a higher-quality solution. We apply our algorithm to the task of learning dynamic textures from image sequences as well as to modeling biosurveillance drug-sales data. The constraint generation approach leads to noticeable improvement in the quality of simulated sequences. We compare our method to those of Lacy and Bernstein [1, 2], with positive results in terms of accuracy, quality of simulated sequences, and efficiency.

****

An online Hebbian learning rule that performs Independent Component Analysis
Claudia Clopath, André Longtin, Wulfram Gerstner
Independent component analysis (ICA) is a powerful method to decouple signals. Most of the algorithms performing ICA do not consider the temporal correlations of the signal, but only higher moments of its amplitude distribution. Moreover, they require some preprocessing of the data (whitening) so as to remove second order correlations. In this paper, we are interested in understanding the neural mechanism responsible for solving ICA. We present an online learning rule that exploits delayed correlations in the input. This rule performs ICA by detecting joint variations in the firing rates of pre- and postsynaptic neurons, similar to a local rate-based Hebbian learning rule.

****

Modeling Natural Sounds with Modulation Cascade Processes
Richard Turner, Maneesh Sahani
Natural sounds are structured on many time-scales. A typical segment of speech, for example, contains features that span four orders of magnitude: Sentences (~1s); phonemes (~0.1s); glottal pulses (~0.01s); and formants (<0.001s). The auditory system uses information from each of these time-scales to solve complicated tasks such as auditory scene analysis. One route toward understanding how audito

ry processing accomplishes this analysis is to build neuroscience-inspired algorithms which solve similar tasks and to compare the properties of these algorithms with properties of auditory processing. There is however a discord: Current machine-audition algorithms largely concentrate on the shorter time-scale structures in sounds, and the longer structures are ignored. The reason for this is two-fold. Firstly, it is a difficult technical problem to construct an algorithm that utilises both sorts of information. Secondly, it is computationally demanding to simultaneously process data both at high resolution (to extract short temporal information) and for long duration (to extract long temporal information). The contribution of this work is to develop a new statistical model for natural sounds that captures structure across a wide range of time-scales, and to provide efficient learning and inference algorithms. We demonstrate the success of this approach on a missing data task.

****

Fast and Scalable Training of Semi-Supervised CRFs with Application to Activity Recognition

Maryam Mahdaviani, Tanzeem Choudhury

We present a new and efficient semi-supervised training method for parameter estimation and feature selection in conditional random fields (CRFs). In real-world applications such as activity recognition, unlabeled sensor traces are relatively easy to obtain whereas labeled examples are expensive and tedious to collect. Furthermore, the ability to automatically select a small subset of discriminatory features from a large pool can be advantageous in terms of computational speed as well as accuracy. In this paper, we introduce the semi-supervised virtual evidence boosting (sVEB) algorithm for training CRFs -- a semi-supervised extension to the recently developed virtual evidence boosting (VEB) method for feature selection and parameter learning. Semi-supervised VEB takes advantage of the unlabeled data via minimum entropy regularization -- the objective function combines the unlabeled conditional entropy with labeled conditional pseudo-likelihood. The sVEB algorithm reduces the overall system cost as well as the human labeling cost required during training, which are both important considerations in building real world inference systems. In a set of experiments on synthetic data and real activity traces collected from wearable sensors, we illustrate that our algorithm benefits from both the use of unlabeled data and automatic feature selection, and outperforms other semi-supervised training approaches.

****

How SVMs can estimate quantiles and the median

Andreas Christmann, Ingo Steinwart

We investigate quantile regression based on the pinball loss and the ■-insensitive loss. For the pinball loss a condition on the data-generating distribution P is given that ensures that the conditional quantiles are approximated with respect to k · k1. This result is then used to derive an oracle inequality for an SVM based on the pinball loss. Moreover, we show that SVMs based on the ■-insensitive loss estimate the conditional median only under certain conditions on P .

****

Random Projections for Manifold Learning

Chinmay Hegde, Michael Wakin, Richard Baraniuk

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-authors prior to requesting a name change in the electronic proceedings.

****

Hippocampal Contributions to Control: The Third Way

Máté Lengyel, Peter Dayan

Recent experimental studies have focused on the specialization of different neural structures for different types of instrumental behavior. Recent theoretical work has provided normative accounts for why there should be more than one control system, and how the output of different controllers can be integrated. Two par- ticlar controllers have been identified, one associated with a forward model and the prefrontal cortex and a second associated with computationally simpler, ha

bit- ual, actor-critic methods and part of the striatum. We argue here for the normative appropriateness of an additional, but so far marginalized control system, associ- ated with episodic memory, and involving the hippocampus and medial temporal cortices. We analyze in depth a class of simple environments to show that episodic control should be useful in a range of cases characterized by complexity and in- ferential noise, and most particularly at the very early stages of learning, long before habitization has set in. We interpret data on the transfer of control from the hippocampus to the striatum in the light of this hypothesis.
****

Rapid Inference on a Novel AND/OR graph for Object Detection, Segmentation and Parsing
Yuanhao Chen, Long Zhu, Chenxi Lin, Hongjiang Zhang, Alan L. Yuille
In this paper we formulate a novel AND/OR graph representation capable of describing the different configurations of deformable articulated objects such as horses. The representation makes use of the summarization principle so that lower level nodes in the graph only pass on summary statistics to the higher level nodes. The probability distributions are invariant to position, orientation, and scale. We develop a novel inference algorithm that combined a bottom-up process for proposing configurations for horses together with a top-down process for refining and validating these proposals. The strategy of surround suppression is applied to ensure that the inference time is polynomial in the size of input data. The algorithm was applied to the tasks of detecting, segmenting and parsing horses. We demonstrate that the algorithm is fast and comparable with the state of the art approaches.
****

Convex Learning with Invariances
Choon Teo, Amir Globerson, Sam Roweis, Alex Smola
Incorporating invariances into a learning algorithm is a common problem in ma- chine learning. We provide a convex formulation which can deal with arbitrary loss functions and arbitrary losses. In addition, it is a drop-in replacement for most optimization algorithms for kernels, including solvers of the SVMStruct family. The advantage of our setting is that it relies on column generation instead of mod- ifying the underlying optimization problem directly.
****

The Noisy-Logical Distribution and its Application to Causal Inference
Alan L. Yuille, Hongjing Lu
We describe a novel noisy-logical distribution for representing the distribution of a binary output variable conditioned on multiple binary input variables. The distribution is represented in terms of noisy-or's and noisy-and-not's of causal features which are conjunctions of the binary inputs. The standard noisy-or and noisy-and-not models, used in causal reasoning and artificial intelligence, are special cases of the noisy-logical distribution. We prove that the noisy-logical distribution is complete in the sense that it can represent all conditional distributions provided a sufficient number of causal factors are used. We illustrate the noisy-logical distribution by showing that it can account for new experimental findings on how humans perform causal reasoning in more complex contexts. Finally, we speculate on the use of the noisy-logical distribution for causal reasoning and artificial intelligence.
****

DIFFRAC: a discriminative and flexible framework for clustering
Francis Bach, Zaïd Harchaoui
We present a novel linear clustering framework (Diffrac) which relies on a linear discriminative cost function and a convex relaxation of a combinatorial optimization problem. The large convex optimization problem is solved through a sequence of lower dimensional singular value decompositions. This framework has several attractive properties: (1) although apparently similar to K-means, it exhibits superior clustering performance than K-means, in particular in terms of robustness to noise. (2) It can be readily extended to non linear clustering if the discriminative cost function is based on positive definite kernels, and can then be seen as an alternative to spectral clustering. (3) Prior information on the par

tition is easily incorporated, leading to state-of-the-art performance for semi-supervised learning, for clustering or classification. We present empirical evaluations of our algorithms on synthetic and real medium-scale datasets.
****

Bundle Methods for Machine Learning
Quoc Le, Alex Smola, S.v.n. Vishwanathan
We present a globally convergent method for regularized risk minimization problems. Our method applies to Support Vector estimation, regression, Gaussian Processes, and any other regularized risk minimization setting which leads to a convex optimization problem. SVMPerf can be shown to be a special case of our approach. In addition to the unified framework we present tight convergence bounds, which show that our algorithm converges in O(1/) steps to precision for general convex problems and in O(log(1/)) steps for continuously differen- tiable problems. We demonstrate in experiments the performance of our approach.
****

Catching Up Faster in Bayesian Model Selection and Model Averaging
Tim Erven, Steven Rooij, Peter Grünwald
Bayesian model averaging, model selection and their approximations such as BIC are generally statistically consistent, but sometimes achieve slower rates of con - vergence than other methods such as AIC and leave-one-out cross-validation. On the other hand, these other methods can be inconsistent. We identify the catch-up phenomenon as a novel explanation for the slow convergence of Bayesian meth-ods. Based on this analysis we define the switch-distribution, a modification of the Bayesian model averaging distribution. We prove that in many situations model selection and prediction based on the switch-distribution is both consistent and achieves optimal convergence rates, thereby resolving the AIC-BIC dilemma. The method is practical; we give an efficient algorithm.
****

Nearest-Neighbor-Based Active Learning for Rare Category Detection
Jingrui He, Jaime Carbonell
Rare category detection is an open challenge for active learning, especially in the de-novo case (no labeled examples), but of significant practical importance for data mining - e.g. detecting new financial transaction fraud patterns, where normal legitimate transactions dominate. This paper develops a new method for detecting an instance of each minority class via an unsupervised local-density-differential sampling strategy. Essentially a variable-scale nearest neighbor process is used to optimize the probability of sampling tightly-grouped minority classes, subject to a local smoothness assumption of the majority class. Results on both synthetic and real data sets are very positive, detecting each minority class with only a frac- tion of the actively sampled points required by random sampling and by Pelleg’s Interleave method, the prior best technique in the sparse literature on this topic.
****

Receptive Fields without Spike-Triggering
Guenther Zeck, Matthias Bethge, Jakob H. Macke
Stimulus selectivity of sensory neurons is often characterized by estimating their receptive field properties such as orientation selectivity. Receptive fields are usually derived from the mean (or covariance) of the spike-triggered stimulus ensemble. This approach treats each spike as an independent message but does not take into account that information might be conveyed through patterns of neural activity that are distributed across space or time. Can we find a concise description for the processing of a whole population of neurons analogous to the receptive field for single neurons? Here, we present a generalization of the linear receptive field which is not bound to be triggered on individual spikes but can be meaningfully linked to distributed response patterns. More precisely, we seek to identify those stimulus features and the corresponding patterns of neural activity that are most reliably coupled. We use an extension of reverse-correlation methods based on canonical correlation analysis. The resulting population receptive fields span the subspace of stimuli that is most informative about the population response. We evaluate our approach using both neuronal models and mult

i-electrode recordings from rabbit retinal ganglion cells. We show how the model can be extended to capture nonlinear stimulus-response relationships using kernel canonical correlation analysis, which makes it possible to test different coding mechanisms. Our technique can also be used to calculate receptive fields from multi-dimensional neural measurements such as those obtained from dynamic imaging methods.

****

Robust Regression with Twinned Gaussian Processes

Andrew Naish-guzman, Sean Holden

We propose a Gaussian process (GP) framework for robust inference in which a GP prior on the mixing weights of a two-component noise model augments the standard process over latent function values. This approach is a generalization of the mixture likelihood used in traditional robust GP regression, and a specialization of the GP mixture models suggested by Tresp (2000) and Rasmussen and Ghahramani (2002). The value of this restriction is in its tractable expectation propagation updates, which allow for faster inference and model selection, and better convergence than the standard mixture. An additional benefit over the latter method lies in our ability to incorporate knowledge of the noise domain to influence predictions, and to recover with the predictive distribution information about the outlier distribution via the gating process. The model has asymptotic complexity equal to that of conventional robust methods, but yields more confident predictions on benchmark problems than classical heavy-tailed models and exhibits improved stability for data with clustered corruptions, for which they fail altogether. We show further how our approach can be used without adjustment for more smoothly heteroscedastic data, and suggest how it could be extended to more general noise models. We also address similarities with the work of Goldberg et al. (1998), and the more recent contributions of Tresp, and Rasmussen and Ghahramani.

****

New Outer Bounds on the Marginal Polytope

David Sontag, Tommi Jaakkola

We give a new class of outer bounds on the marginal polytope, and propose a cutting-plane algorithm for efficiently optimizing over these constraints. When combined with a concave upper bound on the entropy, this gives a new variational inference algorithm for probabilistic inference in discrete Markov Random Fields (MRFs). Valid constraints on the marginal polytope are derived through a series of projections onto the cut polytope. As a result, we obtain tighter upper bounds on the log-partition function. We also show empirically that the approximations of the marginals are significantly more accurate when using the tighter outer bounds. Finally, we demonstrate the advantage of the new constraints for finding the MAP assignment in protein structure prediction.

****

Neural characterization in partially observed populations of spiking neurons

Jonathan Pillow, Peter Latham

Point process encoding models provide powerful statistical methods for under- standing the responses of neurons to sensory stimuli. Although these models have been successfully applied to neurons in the early sensory pathway, they have fared less well capturing the response properties of neurons in deeper brain areas, ow- ing in part to the fact that they do not take into account multiple stages of pro- cessing. Here we introduce a new twist on the point-process modeling approach: we include unobserved as well as observed spiking neurons in a joint encoding model. The resulting model exhibits richer dynamics and more highly nonlinear response properties, making it more powerful and more flexible for fitting neural data. More importantly, it allows us to estimate connectivity patterns among neu- rons (both observed and unobserved), and may provide insight into how networks process sensory input. We formulate the estimation procedure using variational EM and the wake-sleep algorithm, and illustrate the model’s performance using a simulated example network consisting of two coupled neurons.

****

Bayesian Agglomerative Clustering with Coalescents

Yee Teh, Hal Daume III, Daniel M. Roy

We introduce a new Bayesian model for hierarchical clustering based on a prior over trees called Kingman’s coalescent. We develop novel greedy and sequential Monte Carlo inferences which operate in a bottom-up agglomerative fashion. We show experimentally the superiority of our algorithms over the state-of-the-art, and demonstrate our approach in document clustering and phylolinguistics.

****

Distributed Inference for Latent Dirichlet Allocation
David Newman, Padhraic Smyth, Max Welling, Arthur Asuncion

****

Better than least squares: comparison of objective functions for estimating linear-nonlinear models
Tatyana Sharpee

This paper compares a family of methods for characterizing neural feature selec- tivity with natural stimuli in the framework of the linear-nonlinear model. In this model, the neural firing rate is a nonlinear function of a small number of relevant stimulus components. The relevant stimulus dimensions can be found by max- imizing one of the family of objective functions, R´enyi divergences of different orders [1, 2]. We show that maximizing one of them, R´enyi divergence of or- der 2, is equivalent to least-square fitting of the linear-nonlinear model to neural data. Next, we derive reconstruction errors in relevant dimensions found by max- imizing R´enyi divergences of arbitrary order in the asymptotic limit of large spike numbers. We find that the smallest errors are obtained with R´enyi divergence of order 1, also known as Kullback-Leibler divergence. This corresponds to finding relevant dimensions by maximizing mutual information [2]. We numerically test how these optimization schemes perform in the regime of low signal-to-noise ra- tio (small number of spikes and increasing neural noise) for model visual neurons. We find that optimization schemes based on either least square fitting or informa- tion maximization perform well even when number of spikes is small. Information maximization provides slightly, but significantly, better reconstructions than least square fitting. This makes the problem of finding relevant dimensions, together with the problem of lossy compression [3], one of examples where information- theoretic measures are no more data limited than those derived from least squares.

****

Structured Learning with Approximate Inference
Alex Kulesza, Fernando Pereira

In many structured prediction problems, the highest-scoring labeling is hard to compute exactly, leading to the use of approximate inference methods. However, when inference is used in a learning algorithm, a good approximation of the score may not be sufficient. We show in particular that learning can fail even with an approximate inference method with rigorous approximation guarantees. There are two reasons for this. First, approximate methods can effectively reduce the expres- sivity of an underlying model by making it impossible to choose parameters that reliably give good predictions. Second, approximations can respond to parameter changes in such a way that standard learning algorithms are misled. In contrast, we give two positive results in the form of learning bounds for the use of LP-relaxed inference in structured perceptron and empirical risk minimization settings. We argue that without understanding combinations of inference and learning, such as these, that are appropriately compatible, learning performance under approximate inference cannot be guaranteed.

****

On Ranking in Survival Analysis: Bounds on the Concordance Index
Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, Vikas C. Raykar

In this paper, we show that classical survival analysis involving censored data can naturally be cast as a ranking problem. The concordance index (CI), which quantifies the quality of rankings, is the standard performance measure for model \emph{assessment} in survival analysis. In contrast, the standard approach to \emph{learning} the popular proportional hazard (PH) model is based on Cox's parti al likelihood. In this paper we devise two bounds on CI--one of which emerges directly from the properties of PH models--and optimize them \emph{directly}. Our experimental results suggest that both methods perform about equally well, with our new approach giving slightly better results than the Cox's method. We also explain why a method designed to maximize the Cox's partial likelihood also ends up (approximately) maximizing the CI.

****

Competition Adds Complexity

Judy Goldsmith, Martin Mundhenk

It is known that determinining whether a DEC-POMDP, namely, a cooperative partially observable stochastic game (POSG), has a cooperative strategy with positive expected reward is complete for NEXP. It was not known until now how cooperation affected that complexity. We show that, for competitive POSGs, the complexity of determining whether one team has a positive-expected-reward strategy is complete for the class NEXP with an oracle for NP.

****

Classification via Minimum Incremental Coding Length (MICL)

John Wright, Yangyu Tao, Zhouchen Lin, Yi Ma, Heung-yeung Shum

We present a simple new criterion for classification, based on principles from lossy data compression. The criterion assigns a test sample to the class that uses the min- imum number of additional bits to code the test sample, subject to an allowable distortion. We prove asymptotic optimality of this criterion for Gaussian data and analyze its relationships to classical classifiers. Theoretical results provide new insights into relationships among popular classifiers such as MAP and RDA, as well as unsupervised clustering methods based on lossy compression [13]. Mini- mizing the lossy coding length induces a regularization effect which stabilizes the (implicit) density estimate in a small-sample setting. Compression also provides a uniform means of handling classes of varying dimension. This simple classi- fication criterion and its kernel and local versions perform competitively against existing classifiers on both synthetic examples and real imagery data such as hand- written digits and human faces, without requiring domain-specific information.

****

Kernel Measures of Conditional Dependence

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, Bernhard Schölkopf

We propose a new measure of conditional dependence of random variables, based on normalized cross-covariance operators on reproducing kernel Hilbert spaces. Unlike previous kernel dependence measures, the proposed criterion does not de- pend on the choice of kernel in the limit of infinite data, for a wide class of ker- nels. At the same time, it has a straightforward empirical estimate with good convergence behaviour. We discuss the theoretical properties of the measure, and demonstrate its application in experiments.

****

Bayesian Policy Learning with Trans-Dimensional MCMC

Matthew Hoffman, Arnaud Doucet, Nando Freitas, Ajay Jasra

A recently proposed formulation of the stochastic planning and control problem as one of parameter estimation for suitable artificial statistical models has led to the adoption of inference algorithms for this notoriously hard problem. At the algorithmic level, the focus has been on developing Expectation-Maximization (EM) algorithms. In this paper, we begin by making the crucial observation that the stochastic control problem can be reinterpreted as one of trans-dimensional inference. With this new interpretation, we are able to propose a novel reversible jump Markov chain Monte Carlo (MCMC) algorithm that is more efficient than its EM counterparts. Moreover, it enables us to implement full Bayesian policy search, without the need for gradients and with one single Markov chain. The new approach involves sampling directly from a distribution that is proportional to the reward and, consequently, performs better than classic simulations methods in situations where the reward is a rare event.

****

Temporal Difference Updating without a Learning Rate

Marcus Hutter, Shane Legg

We derive an equation for temporal difference learning from statistical principles. Specifically, we start with the variational principle and then bootstrap to produce an updating rule for discounted state value estimates. The resulting equation is similar to the standard equation for temporal difference learning with eligibil- ity traces, so called TD(λ), however it lacks the parameter α that specifies the learning rate. In the place of this free parameter there is now an equation for the learning rate that is specific to each state transition. We experimentally test this new learning rule against TD(λ) and find that it offers superior performance in various settings. Finally, we make some preliminary investigations into how to extend our new temporal difference algorithm to reinforcement learning. To do this we combine our update equation with both Watkins’ Q(λ) and Sarsa(λ) and find that it again offers superior performance without a learning rate parameter.

****

Bayes-Adaptive POMDPs

Stephane Ross, Brahim Chaib-draa, Joelle Pineau

Bayesian Reinforcement Learning has generated substantial interest recently, as it provides an elegant solution to the exploration-exploitation trade-off in reinforce- ment learning. However most investigations of Bayesian reinforcement learning to date focus on the standard Markov Decision Processes (MDPs). Our goal is to extend these ideas to the more general Partially Observable MDP (POMDP) framework, where the state is a hidden variable. To address this problem, we in- troduce a new mathematical model, the Bayes-Adaptive POMDP. This new model allows us to (1) improve knowledge of the POMDP domain through interaction with the environment, and (2) plan optimal sequences of actions which can trade- off between improving the model, identifying the state, and gathering reward. We show how the model can be finitely approximated while preserving the value func- tion. We describe approximations for belief tracking and planning in this model. Empirical results on two domains show that the model estimate and agent’s return improve over time, as the agent learns better model estimates.

****

Regulator Discovery from Gene Expression Time Series of Malaria Parasites: a Hierachical Approach

José Hernández-lobato, Tjeerd Dijkstra, Tom Heskes

We introduce a hierarchical Bayesian model for the discovery of putative regulators from gene expression data only. The hierarchy incorporates the knowledge that there are just a few regulators that by themselves only regulate a handful of genes. This is implemented through a so-called spike-and-slab prior, a mixture of Gaussians with different widths, with mixing weights from a hierarchical Bernoulli model. For efficient inference we implemented expectation propagation. Running the model on a malaria parasite data set, we found four genes with significant homology to transcription factors in an amoebe, one RNA regulator and three genes of unknown function (out of the top ten genes considered).

****

Convex Clustering with Exemplar-Based Models

Danial Lashkari, Polina Golland

Clustering is often formulated as the maximum likelihood estimation of a mixture model that explains the data. The EM algorithm widely used to solve the resulting optimization problem is inherently a gradient-descent method and is sensitive to initialization. The resulting solution is a local optimum in the neighborhood of the initial guess. This sensitivity to initialization presents a significant challenge in clustering large data sets into many clusters. In this paper, we present a dif- ferent approach to approximate mixture fitting for clustering. We introduce an exemplar-based likelihood function that approximates the exact likelihood. This formulation leads to a convex minimization problem and an efficient algorithm with guaranteed convergence to the globally optimal solution. The resulting clus- tering can be thought of as a probabilistic mapping of the data points to the set of exemplars that minimizes the average distance and the information-theoretic cost of mapping. We present experimental results illustrating the pe rformance of our algorithm and its comparison with the conventional approach to mixture model clustering.

****

Learning Bounds for Domain Adaptation

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, Jennifer Wortman

Empirical risk minimization offers well-known learning guarantees when training and test data come from the same domain. In the real world, though, we often wish to adapt a classifier from a source domain with a large amount of training data to different target domain with very little training data. In this work we give uniform convergence bounds for algorithms that minimize a convex combination of source and target empirical risk. The bounds explicitly model the inherent trade-off between training on a large but inaccurate source data set and a small but accurate target training set. Our theory also gives results when we have multiple source domains, each of which may have a different number of instances, and we exhibit cases in which minimizing a non-uniform combination of source risks can achieve much lower target error than standard empirical risk minimization.

****

SpAM: Sparse Additive Models

Han Liu, Larry Wasserman, John Lafferty, Pradeep Ravikumar

We present a new class of models for high-dimensional nonparametric regression and classification called sparse additive models (SpAM). Our methods combine ideas from sparse linear modeling and additive nonparametric regression. We de- rive a method for fitting the models that is effective even when the number of covariates is larger than the sample size. A statistical analysis of the properties of SpAM is given together with empirical results on synthetic and real data, showing that SpAM can be effective in fitting sparse nonparametric models in high dimensional data.

****

Bayesian Inference for Spiking Neuron Models with a Sparsity Prior

Sebastian Gerwinn, Matthias Bethge, Jakob H. Macke, Matthias Seeger

Generalized linear models are the most commonly used tools to describe the stimulus selectivity of sensory neurons. Here we present a Bayesian treatment of such models. Using the expectation propagation algorithm, we are able to approximate the full posterior distribution over all weights. In addition, we use a Laplacian prior to favor sparse solutions. Therefore, stimulus features that do not critically influence neural activity will be assigned zero weights and thus be effectively excluded by the model. This feature selection mechanism facilitates both the in- terpretation of the neuron model as well as its predictive abilities. The posterior distribution can be used to obtain confidence intervals which makes it possible to assess the statistical significance of the solution. In neural data analysis, the available amount of experimental measurements is often limited whereas the pa- rameter space is large. In such a situation, both regularization by a sparsity prior and uncertainty estimates for the model parameters are essential. We apply our method to multi-electrode recordings of retinal ganglion cells and use our uncer- tainty estimate to test the statistical significance of functional couplings between neurons. Furthermore we used the sparsity of the Laplace prior to select those filters from a spike-triggered covariance analysis that are most informative about the neural response.

****

Unconstrained On-line Handwriting Recognition with Recurrent Neural Networks

Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, Santiago Fernández

On-line handwriting recognition is unusual among sequence labelling tasks in that the underlying generator of the observed data, i.e. the movement of the pen, is recorded directly. However, the raw data can be difficult to interpret because each letter is spread over many pen locations. As a consequence, sophisticated pre-processing is required to obtain inputs suitable for conventional sequence labelling algorithms, such as HMMs. In this paper we describe a system capable of directly transcribing raw on-line handwriting data. The system consists of a recurrent neural network trained for sequence labelling, combined with a probabilistic language model. In experiments on an unconstrained on-line database, we rec ord excellent results using either raw or pre-processed data, well outperforming a benchmark HMM in both cases.

****

The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information

John Langford, Tong Zhang

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-authors prior to requesting a name change in the electronic proceedings.

****

Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes

Geoffrey E. Hinton, Russ R. Salakhutdinov

We show how to use unlabeled data and a deep belief net (DBN) to learn a good covariance kernel for a Gaussian process. We first learn a deep generative model of the unlabeled data using the fast, greedy algorithm introduced by Hinton et.al. If the data is high-dimensional and highly-structured, a Gaussian kernel applied to the top layer of features in the DBN works much better than a similar kernel applied to the raw input. Performance at both regression and classification can then be further improved by using backpropagation through the DBN to discriminatively fine-tune the covariance kernel.

****

Kernels on Attributed Pointsets with Applications

Mehul Parsana, Sourangshu Bhattacharya, Chiru Bhattacharya, K. Ramakrishnan

This paper introduces kernels on attributed pointsets, which are sets of vectors embedded in an euclidean space. The embedding gives the notion of neighborhood, which is used to define positive semidefinite kernels on pointsets. Two novel kernels on neighborhoods are proposed, one evaluating the attribute similarity and the other evaluating shape similarity. Shape similarity function is motivated from spectral graph matching techniques. The kernels are tested on three real life applications: face recognition, photo album tagging, and shot annotation in video sequences, with encouraging results.

****

Testing for Homogeneity with Kernel Fisher Discriminant Analysis

Moulines Eric, Francis Bach, Zaïd Harchaoui

We propose to test for the homogeneity of two samples by using Kernel Fisher discriminant Analysis. This provides us with a consistent nonparametric test statistic, for which we derive the asymptotic distribution under the null hypothesis. We give experimental evidence of the relevance of our method on both artificial and real datasets.

****

Sparse deep belief net model for visual area V2

Honglak Lee, Chaitanya Ekanadham, Andrew Ng

Motivated in part by the hierarchical organization of cortex, a number of algorithms have recently been proposed that try to learn hierarchical, or deep,'' structure from unlabeled data. While several authors have formally or informally compared their algorithms to computations performed in visual area V1 (and the cochlea), little attempt has been made thus far to evaluate these algorithms in terms of their fidelity for mimicking computations at deeper levels in the cortical hierarchy. This paper presents an unsupervised learning model that faithfully mimics certain properties of visual area V2. Specifically, we develop a sparse variant of the deep belief networks of Hinton et al. (2006). We learn two layers of nodes in the network, and demonstrate that the first layer, similar to prior work on sparse coding and ICA, results in localized, oriented, edge filters, similar to the Gabor functions known to model V1 cell receptive fields. Further, the second layer in our model encodes correlations of the first layer responses in the data. Specifically, it picks up both collinear (contour'') features as well as corners and junctions. More interestingly, in a quantitative comparison, the encoding of these more complex ``corner'' features matches well with the results from the Ito & Komatsu's study of biological V2 responses. This suggests that our sparse variant of deep belief networks holds promise for modeling more higher-order features.

****

Second Order Bilinear Discriminant Analysis for single trial EEG analysis

Christoforos Christoforou, Paul Sajda, Lucas Parra

Traditional analysis methods for single-trial classification of electro-encephalography (EEG) focus on two types of paradigms: phase locked methods, in which the amplitude of the signal is used as the feature for classification, i.e. event related potentials; and second order methods, in which the feature of interest is the power of the signal, i.e event related (de)synchronization. The process of deciding which paradigm to use is ad hoc and is driven by knowledge of neurological findings. Here we propose a unified method in which the algorithm learns the best first and second order spatial and temporal features for classification of EEG based on a bilinear model. The efficiency of the method is demonstrated in simulated and real EEG from a benchmark data set for Brain Computer Interface.

****

Convex Relaxations of Latent Variable Training

Yuhong Guo, Dale Schuurmans

We investigate a new, convex relaxation of an expectation-maximization (EM) variant that approximates a standard objective while eliminating local minima. First, a cautionary result is presented, showing that any convex relaxation of EM over hidden variables must give trivial results if any dependence on the missing values is retained. Although this appears to be a strong negative outcome, we then demonstrate how the problem can be bypassed by using equivalence relations instead of value assignments over hidden variables. In particular, we develop new algorithms for estimating exponential conditional models that only require equivalence relation information over the variable values. This reformulation leads to an exact expression for EM variants in a wide range of problems. We then develop a semidefinite relaxation that yields global training by eliminating local minima.

****

A configurable analog VLSI neural network with spiking neurons and self-regulating plastic synapses

Massimiliano Giulioni, Mario Pannunzi, Davide Badoni, Vittorio Dante, Paolo Giudice

We summarize the implementation of an analog VLSI chip hosting a network of 32 integrate-and-fire (IF) neurons with spike-frequency adaptation and 2,048 Hebbian plastic bistable spike-driven stochastic synapses endowed with a self-regulating mechanism which stops unnecessary synaptic changes. The synaptic matrix can be flexibly configured and provides both recurrent and AER-based connectivity with external, AER compliant devices. We demonstrate the ability of the network to efficiently classify overlapping patterns, thanks to the self-regulating mechanism.

****

The discriminant center-surround hypothesis for bottom-up saliency

Dashan Gao, Vijay Mahadevan, Nuno Vasconcelos

The classical hypothesis, that bottom-up saliency is a center-surround process, is combined with a more recent hypothesis that all saliency decisions are optimal in a decision-theoretic sense. The combined hypothesis is denoted as discriminant center-surround saliency, and the corresponding optimal saliency architecture is derived. This architecture equates the saliency of each image location to the discriminant power of a set of features with respect to the classification problem that opposes stimuli at center and surround, at that location. It is shown that the resulting saliency detector makes accurate quantitative predictions for various aspects of the psychophysics of human saliency, including non-linear properties beyond the reach of previous saliency models. Furthermore, it is shown that discriminant center-surround saliency can be easily generalized to various stimulus modalities (such as color, orientation and motion), and provides optimal solutions for many other saliency problems of interest for computer vision. Optimal solutions, under this hypothesis, are derived for a number of the former (including static natural images, dense motion fields, and even dynamic textures

), and applied to a number of the latter (the prediction of human eye fixations, motion-based saliency in the presence of ego-motion, and motion-based saliency in the presence of highly dynamic backgrounds). In result, discriminant saliency is shown to predict eye fixations better than previous models, and produce background subtraction algorithms that outperform the state-of-the-art in computer vision.

****

Statistical Analysis of Semi-Supervised Regression

Larry Wasserman, John Lafferty

Semi-supervised methods use unlabeled data in addition to labeled data to con- struct predictors. While existing semi-supervised methods have shown some promising empirical performance, their development has been based largely based on heuristics. In this paper we study semi-supervised learning from the viewpoint of minimax theory. Our first result shows that some common methods based on regularization using graph Laplacians do not lead to faster minimax rates of con- vergence. Thus, the estimators that use the unlabeled data do not have smaller risk than the estimators that use only labeled data. We then develop several new approaches that provably lead to improved performance. The statistical tools of minimax analysis are thus used to offer some new perspective on the problem of semi-supervised learning.

****

Hierarchical Apprenticeship Learning with Application to Quadruped Locomotion

J. Kolter, Pieter Abbeel, Andrew Ng

We consider apprenticeship learning—learning from expert demonstrations—in the setting of large, complex domains. Past work in apprenticeship learning requires that the expert demonstrate complete trajectories through the domain. However, in many problems even an expert has difficulty controlling the system, which makes this approach infeasible. For example, consider the task of teach- ing a quadruped robot to navigate over extreme terrain; demonstrating an optimal policy (i.e., an optimal set of foot locations over the entire terrain) is a highly non-trivial task, even for an expert. In this paper we propose a method for hier- archical apprenticeship learning, which allows the algorithm to accept isolated advice at different hierarchical levels of the control task. This type of advice is often feasible for experts to give, even if the expert is unable to demonstrate com- plete trajectories. This allows us to extend the apprenticeship learning paradigm to much larger, more challenging domains. In particular, in this paper we apply the hierarchical apprenticeship learning algorithm to the task of quadruped loco- motion over extreme terrain, and achieve, to the best of our knowledge, results superior to any previously published work.

****

Colored Maximum Variance Unfolding

Le Song, Arthur Gretton, Karsten Borgwardt, Alex Smola

Maximum variance unfolding (MVU) is an effective heuristic for dimensionality reduction. It produces a low-dimensional representation of the data by maximiz- ing the variance of their embeddings while preserving the local distances of the original data. We show that MVU also optimizes a statistical dependence measure which aims to retain the identity of individual observations under the distance-preserving constraints. This general view allows us to design “colored” variants of MVU, which produce low-dimensional representations for a given task, e.g. subject to class labels or other side information.

****

Adaptive Embedded Subgraph Algorithms using Walk-Sum Analysis

Venkat Chandrasekaran, Alan Willsky, Jason Johnson

We consider the estimation problem in Gaussian graphical models with arbitrary structure. We analyze the Embedded Trees algorithm, which solves a sequence of problems on tractable subgraphs thereby leading to the solution of the estimation problem on an intractable graph. Our analysis is based on the recently developed walk-sum interpretation of Gaussian estimation. We show that non-stationary iterations of the Embedded Trees algorithm using any sequence of subgraphs converge in walk-summable models. Based on walk-sum calculations, we develop adaptive me thods that optimize the choice of subgraphs used at each iteration with a view to achieving maximum reduction in error. These adaptive procedures provide a significant speedup in convergence over stationary iterative methods, and also appear to converge in a larger class of models.

****

Ultrafast Monte Carlo for Statistical Summations

Charles Isbell, Michael Holmes, Alexander Gray

Machine learning contains many computational bottlenecks in the form of nested summations over datasets. Kernel estimators and other methods are burdened by these expensive computations. Exact evaluation is typically O(n2) or higher, which severely limits application to large datasets. We present a multi-stage stratified Monte Carlo method for approximating such summations with probabilistic relative error control. The essential idea is fast approximation by sampling in trees. This method differs from many previous scalability techniques (such as standard multi-tree methods) in that its error is stochastic, but we derive conditions for error control and demonstrate that they work. Further, we give a theoretical sample complexity for the method that is independent of dataset size, and show that this appears to hold in experiments, where speedups reach as high as 1014, many orders of magnitude beyond the previous state of the art.

****

Inferring Neural Firing Rates from Spike Trains Using Gaussian Processes

John P. Cunningham, Byron M. Yu, Krishna V. Shenoy, Maneesh Sahani

Neural spike trains present challenges to analytical efforts due to their noisy, spiking nature. Many studies of neuroscienti(cid:2)c and neural prosthetic importance rely on a smoothed, denoised estimate of the spike train’s underlying (cid:2)ring rate. Current techniques to (cid:2)nd time-varying (cid:2)ring rates require ad hoc choices of parameters, offer no con(cid:2)dence intervals on their estimates, and can obscure potentially important single trial variability. We present a new method, based on a Gaussian Process prior, for inferring probabilistically optimal estimates of (cid:2)ring rate functions underlying single or multiple neural spike trains. We test the performance of the method on simulated data and experimentally gathered neural spike trains, and we demonstrate improvements over conventional estimators.

****

People Tracking with the Laplacian Eigenmaps Latent Variable Model

Zhengdong Lu, Cristian Sminchisescu, Miguel Carreira-Perpiñán

Reliably recovering 3D human pose from monocular video requires constraints that bias the estimates towards typical human poses and motions. We define priors for people tracking using a Laplacian Eigenmaps Latent Variable Model (LELVM). LELVM is a probabilistic dimensionality reduction model that naturally combines the advantages of latent variable models---definining a multimodal probability density for latent and observed variables, and globally differentiable nonlinear mappings for reconstruction and dimensionality reduction---with those of spectral manifold learning methods---no local optima, ability to unfold highly nonlinear manifolds, and good practical scaling to latent spaces of high dimension. LELVM is computationally efficient, simple to learn from sparse training data, and compatible with standard probabilistic trackers such as particle filters. We analyze the performance of a LELVM-based probabilistic sigma point mixture tracker in several real and synthetic human motion sequences and demonstrate that LELVM provides sufficient constraints for robust operation in the presence of missing, noisy and ambiguous image measurements.

****

The Distribution Family of Similarity Distances

Gertjan Burghouts, Arnold Smeulders, Jan-mark Geusebroek

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-authors prior to requesting a name change in the electronic proceedings.

****

Congruence between model and human attention reveals unique signatures of critic

al visual events
Robert Peters, Laurent Itti
Current computational models of bottom-up and top-down components of atten- tion are predictive of eye movements across a range of stimuli and of simple, fixed visual tasks (such as visual search for a target among distractors). How- ever, to date there exists no computational framework which can reliably mimic human gaze behavior in more complex environments and tasks, such as driving a vehicle through traffic. Here, we develop a hybrid computational/behavioral framework, combining simple models for bottom-up salience and top-down rel- evance, and looking for changes in the predictive power of these components at different critical event times during 4.7 hours (500,000 video frames) of observers playing car racing and flight combat video games. This approach is motivated by our observation that the predictive strengths of the salience and relevance mod- els exhibit reliable temporal signatures during critical event windows in the task sequence—for example, when the game player directly engages an enemy plane in a flight combat game, the predictive strength of the salience model increases significantly, while that of the relevance model decreases significantly. Our new framework combines these temporal signatures to implement several event detec- tors. Critically, we find that an event detector based on fused behavioral and stim- ulus information (in the form of the model’s predictive strength) is much stronger than detectors based on behavioral information alone (eye position) or image in- formation alone (model prediction maps). This approach to event detection, based on eye tracking combined with computational models applied to the visual input, may have useful applications as a less-invasive alternative to other event detection approaches based on neural signatures derived from EEG or fMRI recordings.
****
Multi-task Gaussian Process Prediction
Edwin V. Bonilla, Kian Chai, Christopher Williams
In this paper we investigate multi-task learning in the context of Gaussian Pro- cesses (GP). We propose a model that learns a shared covariance function on input-dependent features and a “free-form” covariance matrix over tasks. This al- lows for good flexibility when modelling inter-task dependencies while avoiding the need for large amounts of data for training. We show that under the assump- tion of noise-free observations and a block design, predictions for a given task only depend on its target values and therefore a cancellation of inter-task trans- fer occurs. We evaluate the benefits of our model on two practical applications: a compiler performance prediction problem and an exam score prediction task. Additionally, we make use of GP approximations and properties of our model in order to provide scalability to large data sets.
****
Multi-Task Learning via Conic Programming
Tsuyoshi Kato, Hisashi Kashima, Masashi Sugiyama, Kiyoshi Asai
When we have several related tasks, solving them simultaneously is shown to be more effective than solving them individually. This approach is called multi-task learning (MTL) and has been studied extensively. Existing approaches to MTL often treat all the tasks as \emph{uniformly related to each other and the relatedness of the tasks is controlled globally. For this reason, the existing methods can lead to undesired solutions when some tasks are not highly related to each other, and some pairs of related tasks can have significantly different solutions. In this paper, we propose a novel MTL algorithm that can overcome these problems. Our method makes use of a task network, which describes the relation structure among tasks. This allows us to deal with intricate relation structures in a systematic way. Furthermore, we control the relatedness of the tasks locally, so all pairs of related tasks are guaranteed to have similar solutions. We apply the above idea to support vector machines (SVMs) and show that the optimization problem can be cast as a second order cone program, which is convex and can be solved efficiently. The usefulness of our approach is demonstrated through simulations with protein super-family classification and ordinal regression problems.
****
Incremental Natural Actor-Critic Algorithms

Shalabh Bhatnagar, Mohammad Ghavamzadeh, Mark Lee, Richard S. Sutton

We present four new reinforcement learning algorithms based on actor-critic and natural-gradient ideas, and provide their convergence proofs. Actor-critic reinforcement learning methods are online approximations to policy iteration in which the value-function parameters are estimated using temporal difference learning and the policy parameters are updated by stochastic gradient descent. Methods based on policy gradients in this way are of special interest because of their com- patibility with function approximation methods, which are needed to handle large or in(cid:2)nite state spaces. The use of temporal difference learning in this way is of interest because in many applications it dramatically reduces the variance of the gradient estimates. The use of the natural gradient is of interest because it can produce better conditioned parameterizations and has been shown to further re- duce variance in some cases. Our results extend prior two-timescale convergence results for actor-critic methods by Konda and Tsitsiklis by using temporal differ- ence learning in the actor and by incorporating natural gradients, and they extend prior empirical studies of natural actor-critic methods by Peters, Vijayakumar and Schaal by providing the (cid:2)rst convergence proofs and the (cid:2)rst fully incremental algorithms.

****

Collective Inference on Markov Models for Modeling Bird Migration

M.a. Elmohamed, Dexter Kozen, Daniel R. Sheldon

We investigate a family of inference problems on Markov models, where many sample paths are drawn from a Markov chain and partial information is revealed to an observer who attempts to reconstruct the sample paths. We present algo- rithms and hardness results for several variants of this problem which arise by re- vealing different information to the observer and imposing different requirements for the reconstruction of sample paths. Our algorithms are analogous to the clas- sical Viterbi algorithm for Hidden Markov Models, which finds the single most probable sample path given a sequence of observations. Our work is motivated by an important application in ecology: inferring bird migration paths from a large database of observations.

****

EEG-Based Brain-Computer Interaction: Improved Accuracy by Automatic Single-Trial Error Detection

Pierre Ferrez, José Millán

Brain-computer interfaces (BCIs), as any other interaction modality based on physiological signals and body channels (e.g., muscular activity, speech and gestures), are prone to errors in the recognition of subject's intent. An elegant approach to improve the accuracy of BCIs consists in a verification procedure directly based on the presence of error-related potentials (ErrP) in the EEG recorded right after the occurrence of an error. Six healthy volunteer subjects with no prior BCI experience participated in a new human-robot interaction experiment where they were asked to mentally move a cursor towards a target that can be reached within a few steps using motor imagination. This experiment confirms the previously reported presence of a new kind of ErrP. These Interaction ErrP" exhibit a first sharp negative peak followed by a positive peak and a second broader negative peak (~290, ~350 and ~470 ms after the feedback, respectively). But in order to exploit these ErrP we need to detect them in each single trial using a short window following the feedback associated to the response of the classifier embedded in the BCI. We have achieved an average recognition rate of correct and erroneous single trials of 81.8% and 76.2%, respectively. Furthermore, we have achieved an average recognition rate of the subject's intent while trying to mentally drive the cursor of 73.1%. These results show that it's possible to simultaneously extract useful information for mental control to operate a brain-actuated device as well as cognitive states such as error potentials to improve the quality of the brain-computer interaction. Finally, using a well-known inverse model (sLORETA), we show that the main focus of activity at the occurrence of the ErrP are, as expected, in the pre-supplementary motor area and in the anterior cingulate cortex."

****

Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing

Benjamin Blankertz, Motoaki Kawanabe, Ryota Tomioka, Friederike Hohlefeld, Klaus-Robert Müller, Vadim Nikulin

Brain-Computer Interfaces can suffer from a large variance of the subject conditions within and across sessions. For example vigilance fluctuations in the individual, variable task involvement, workload etc. alter the characteristics of EEG signals and thus challenge a stable BCI operation. In the present work we aim to define features based on a variant of the common spatial patterns (CSP) algorithm that are constructed invariant with respect to such nonstationarities. We enforce invariance properties by adding terms to the denominator of a Rayleigh coefficient representation of CSP such as disturbance covariance matrices from fluctuations in visual processing. In this manner physiological prior knowledge can be used to shape the classification engine for BCI. As a proof of concept we present a BCI classifier that is robust to changes in the level of parietal a -activity. In other words, the EEG decoding still works when there are lapses in vigilance.

****

The Infinite Gamma-Poisson Feature Model

Michalis Titsias

We address the problem of factorial learning which associates a set of latent causes or features with the observed data. Factorial models usually assume that each feature has a single occurrence in a given data point. However, there are data such as images where latent features have multiple occurrences, e.g. a visual object class can have multiple instances shown in the same image. To deal with such cases, we present a probability model over non-negative integer valued matrices with possibly unbounded number of columns. This model can play the role of the prior in an nonparametric Bayesian learning scenario where both the latent features and the number of their occurrences are unknown. We use this prior together with a likelihood model for unsupervised learning from images using a Markov Chain Monte Carlo inference algorithm.

****

A Unified Near-Optimal Estimator For Dimension Reduction in l_α ($0<\alpha\leq 2$) Using Stable Random Projections

Ping Li, Trevor Hastie

Many tasks (e.g., clustering) in machine learning only require the lα distances in- stead of the original data. For dimension reductions in the lα norm ($0 < \alpha \leq 2$), the method of stable random projections can efficiently compute the lα distances in massive datasets (e.g., the Web or massive data streams) in one pass of the data. The estimation task for stable random projections has been an interesting topic. We propose a simple estimator based on the fractional power of the samples (pro- jected data), which is surprisingly near-optimal in terms of the asymptotic vari- ance. In fact, it achieves the Cram´er-Rao bound when $\alpha = 2$ and $\alpha = 0+$. This new result will be useful when applying stable random projections to distance- based clustering, classifications, kernels, massive data streams etc.

****

Continuous Time Particle Filtering for fMRI

Lawrence Murray, Amos J. Storkey

We construct a biologically motivated stochastic differential model of the neural and hemodynamic activity underlying the observed Blood Oxygen Level Dependent (BOLD) signal in Functional Magnetic Resonance Imaging (fMRI). The model poses a difficult parameter estimation problem, both theoretically due to the nonlinearity and divergence of the differential system, and computationally due to its time and space complexity. We adapt a particle filter and smoother to the task, and discuss some of the practical approaches used to tackle the difficulties, including use of sparse matrices and parallelisation. Results demonstrate the tractability of the approach in its application to an effective connectivity study.

****

Computing Robust Counter-Strategies

Michael Johanson, Martin Zinkevich, Michael Bowling

Adaptation to other initially unknown agents often requires computing an effective counter-strategy. In the Bayesian paradigm, one must find a good counter-strategy to the inferred posterior of the other agents' behavior. In the experts paradigm, one may want to choose experts that are good counter-strategies to the other agents' expected behavior. In this paper we introduce a technique for computing robust counter-strategies for adaptation in multiagent scenarios under a variety of paradigms. The strategies can take advantage of a suspected tendency in the decisions of the other agents, while bounding the worst-case performance when the tendency is not observed. The technique involves solving a modified game, and therefore can make use of recently developed algorithms for solving very large extensive games. We demonstrate the effectiveness of the technique in two-player Texas Hold'em. We show that the computed poker strategies are substantially more robust than best response counter-strategies, while still exploiting a suspected tendency. We also compose the generated strategies in an experts algorithm showing a dramatic improvement in performance over using simple best responses.

****

Random Sampling of States in Dynamic Programming

Chris Atkeson, Benjamin Stephens

We combine two threads of research on approximate dynamic programming: random sampling of states and using local trajectory optimizers to globally optimize a policy and associated value function. This combination allows us to replace a dense multidimensional grid with a much sparser adaptive sampling of states. Our focus is on finding steady state policies for the deterministic time invariant discrete time control problems with continuous states and actions often found in robotics. In this paper we show that we can now solve problems we couldn't solve previously with regular grid-based approaches.

****

Predicting human gaze using low-level saliency combined with face detection

Moran Cerf, Jonathan Harel, Wolfgang Einhaeuser, Christof Koch

Under natural viewing conditions, human observers shift their gaze to allocate processing resources to subsets of the visual input. Many computational models have aimed at predicting such voluntary attentional shifts. Although the importance of high level stimulus properties (higher order statistics, semantics) stands undisputed, most models are based on low-level features of the input alone. In this study we recorded eye-movements of human observers while they viewed photographs of natural scenes. About two thirds of the stimuli contained at least one person. We demonstrate that a combined model of face detection and low-level saliency clearly outperforms a low-level model in predicting locations humans fixate. This is reflected in our finding fact that observes, even when not instructed to look for anything particular, fixate on a face with a probability of over 80% within their first two fixations (500ms). Remarkably, the model's predictive performance in images that do not contain faces is not impaired by spurious face detector responses, which is suggestive of a bottom-up mechanism for face detection. In summary, we provide a novel computational approach which combines high level object knowledge (in our case: face locations) with low-level features to successfully predict the allocation of attentional resources.

****

Local Algorithms for Approximate Inference in Minor-Excluded Graphs

Kyomin Jung, Devavrat Shah

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-authors prior to requesting a name change in the electronic proceedings.

****

Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization

XuanLong Nguyen, Martin J. Wainwright, Michael Jordan

We develop and analyze an algorithm for nonparametric estimation of divergence functionals and the density ratio of two probability distributions. Our method is based on a variational characterization of f-divergences, which turns the estim a- tion into a penalized convex risk minimization problem. We present a derivation of our kernel-based estimation algorithm and an analysis of convergence rates for the estimator. Our simulation results demonstrate the convergence behavior of the method, which compares favorably with existing methods in the literature.

****

Learning with Tree-Averaged Densities and Distributions
Sergey Kirshner

We utilize the ensemble of trees framework, a tractable mixture over super- exponential number of tree-structured distributions [1], to develop a new model for multivariate density estimation. The model is based on a construction of tree- structured copulas – multivariate distributions with uniform on [0, 1] marginals. By averaging over all possible tree structures, the new model can approximate distributions with complex variable dependencies. We propose an EM algorithm to estimate the parameters for these tree-averaged models for both the real-valued and the categorical case. Based on the tree-averaged framework, we propose a new model for joint precipitation amounts data on networks of rain stations.

****

Variational inference for Markov jump processes
Manfred Opper, Guido Sanguinetti

Markov jump processes play an important role in a large number of application domains. However, realistic systems are analytically intractable and they have traditionally been analysed using simulation based techniques, which do not provide a framework for statistical inference. We propose a mean field approximation to perform posterior inference and parameter estimation. The approximation allows a practical solution to the inference problem, {while still retaining a good degree of accuracy.} We illustrate our approach on two biologically motivated systems.

****

Expectation Maximization and Posterior Constraints
Kuzman Ganchev, Ben Taskar, João Gama

The expectation maximization (EM) algorithm is a widely used maximum likelihood estimation procedure for statistical models when the values of some of the variables in the model are not observed. Very often, however, our aim is primarily to find a model that assigns values to the latent variables that have intended meaning for our data and maximizing expected likelihood only sometimes accomplishes this. Unfortunately, it is typically difficult to add even simple a-priori information about latent variables in graphical models without making the models overly complex or intractable. In this paper, we present an efficient, principled way to inject rich constraints on the posteriors of latent variables into the EM algorithm. Our method can be used to learn tractable graphical models that satisfy additional, otherwise intractable constraints. Focusing on clustering and the alignment problem for statistical machine translation, we show that simple, intuitive posterior constraints can greatly improve the performance over standard baselines and be competitive with more complex, intractable models.

****

Anytime Induction of Cost-sensitive Trees
Saher Esmeir, Shaul Markovitch

Machine learning techniques are increasingly being used to produce a wide-range of classifiers for complex real-world applications that involve nonuniform testing costs and misclassification costs. As the complexity of these applications grows, the management of resources during the learning and classification processes be- comes a challenging task. In this work we introduce ACT (Anytime Cost-sensitive Trees), a novel framework for operating in such environments. ACT is an anytime algorithm that allows trading computation time for lower classification costs. It builds a tree top-down and exploits additional time resources to obtain better esti- mations for the utility of the different candidate splits. Using sampling techniques ACT approximates for each candidate split the cost of the subtree under it and fa- vors the one with a minimal cost. Due to its stochastic nature ACT is expected to be able to escape local minima, into which greedy methods may be trapped. Ex- periments with a variety of datasets were conducted to compare the performance of ACT to that of the state of the art cost-sensitive tree learners. The results show that for most domains ACT produces trees of significantly lower costs. ACT is also shown to exhibit good anytime behavior with diminishing returns.

****

Optimal ROC Curve for a Combination of Classifiers

Marco Barreno, Alvaro Cardenas, J. D. Tygar

We present a new analysis for the combination of binary classifiers. We propose a theoretical framework based on the Neyman-Pearson lemma to analyze combinations of classifiers. In particular, we give a method for finding the optimal decision rule for a combination of classifiers and prove that it has the optimal ROC curve. We also show how our method generalizes and improves on previous work on combining classifiers and generating ROC curves.

****

Modeling homophily and stochastic equivalence in symmetric relational data

Peter Hoff

This article discusses a latent variable model for inference and prediction of symmetric relational data. The model, based on the idea of the eigenvalue decomposition, represents the relationship between two nodes as the weighted inner-product of node-specific vectors of latent characteristics. This ``eigenmodel'' generalizes other popular latent variable models, such as latent class and distance models: It is shown mathematically that any latent class or distance model has a representation as an eigenmodel, but not vice-versa. The practical implications of this are examined in the context of three real datasets, for which the eigenmodel has as good or better out-of-sample predictive performance than the other two models.

****

On Sparsity and Overcompleteness in Image Models

Pietro Berkes, Richard Turner, Maneesh Sahani

Computational models of visual cortex, and in particular those based on sparse coding, have enjoyed much recent attention. Despite this currency, the question of how sparse or how over-complete a sparse representation should be, has gone without principled answer. Here, we use Bayesian model-selection methods to address these questions for a sparse-coding model based on a Student-t prior. Having validated our methods on toy data, we find that natural images are indeed best modelled by extremely sparse distributions; although for the Student-t prior, the associated optimal basis size is only modestly overcomplete.

****

A Probabilistic Approach to Language Change

Alexandre Bouchard-côté, Percy S. Liang, Dan Klein, Thomas Griffiths

We present a probabilistic approach to language change in which word forms are represented by phoneme sequences that undergo stochastic edits along the branches of a phylogenetic tree. Our framework combines the advantages of the classical comparative method with the robustness of corpus-based probabilistic models. We use this framework to explore the consequences of two different schemes for defining probabilistic models of phonological change, evaluating these schemes using the reconstruction of ancient word forms in Romance languages. The result is an efficient inference procedure for automatically inferring ancient word forms from modern languages, which can be generalized to support inferences about linguistic phylogenies.

****

Learning the 2-D Topology of Images

Nicolas Roux, Yoshua Bengio, Pascal Lamblin, Marc Joliveau, Balázs Kégl

We study the following question: is the two-dimensional structure of images a very strong prior or is it something that can be learned with a few examples of natural images? If someone gave us a learning task involving images for which the two-dimensional topology of pixels was not known, could we discover it automatically and exploit it? For example suppose that the pixels had been permuted in a fixed but unknown way, could we recover the relative two-dimensional location of pixels on images? The surprising result presented here is that not only the ans wer is yes but that about as few as a thousand images are enough to approximately recover the relative locations of about a thousand pixels. This is achieved using a manifold learning algorithm applied to pixels associated with a measure of distributional similarity between pixel intensities. We compare different topology-extraction approaches and show how having the two-dimensional topology can be exploited.

****

A Bayesian LDA-based model for semi-supervised part-of-speech tagging

Kristina Toutanova, Mark Johnson

We present a novel Bayesian model for semi-supervised part-of-speech tagging. Our model extends the Latent Dirichlet Allocation model and incorporates the intuition that words’ distributions over tags, p(t|w), are sparse. In addition we introduce a model for determining the set of possible tags of a word which captures important dependencies in the ambiguity classes of words. Our model outper- forms the best previously proposed model for this task on a standard dataset.

****

Cluster Stability for Finite Samples

Ohad Shamir, Naftali Tishby

Over the past few years, the notion of stability in data clustering has received growing attention as a cluster validation criterion in a sample-based framework. However, recent work has shown that as the sample size increases, any clustering model will usually become asymptotically stable. This led to the conclusion that stability is lacking as a theoretical and practical tool. The discrepancy between this conclusion and the success of stability in practice has remained an open ques- tion, which we attempt to address. Our theoretical approach is that stability, as used by cluster validation algorithms, is similar in certain respects to measures of generalization in a model-selection framework. In such cases, the model cho- sen governs the convergence rate of generalization bounds. By arguing that these rates are more important than the sample size, we are led to the prediction that stability-based cluster validation algorithms should not degrade with increasing sample size, despite the asymptotic universal stability. This prediction is substan- tiated by a theoretical analysis as well as some empirical results. We conclude that stability remains a meaningful cluster validation criterion over finite samples.

****

Variational Inference for Diffusion Processes

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, John Shawe-taylor

Diffusion processes are a family of continuous-time continuous-state stochastic processes that are in general only partially observed. The joint estimation of the forcing parameters and the system noise (volatility) in these dynamical systems is a crucial, but non-trivial task, especially when the system is nonlinear and multi-modal. We propose a variational treatment of diffusion processes, which allows us to estimate these parameters by simple gradient techniques and which is computationally less demanding than most MCMC approaches. Furthermore, our parameter inference scheme does not break down when the time step gets smaller, unlike most current approaches. Finally, we show how a cheap estimate of the posterior over the parameters can be constructed based on the variational free energy.

****

Augmented Functional Time Series Representation and Forecasting with Gaussian Processes

Nicolas Chapados, Yoshua Bengio

We introduce a functional representation of time series which allows forecasts to be performed over an unspecified horizon with progressively-revealed informa- tion sets. By virtue of using Gaussian processes, a complete covariance matrix between forecasts at several time-steps is available. This information is put to use in an application to actively trade price spreads between commodity futures con- tracts. The approach delivers impressive out-of-sample risk-adjusted returns after transaction costs on a portfolio of 30 spreads.

****

Sparse Overcomplete Latent Variable Decomposition of Counts Data
Madhusudana Shashanka, Bhiksha Raj, Paris Smaragdis

An important problem in many fields is the analysis of counts data to extract meaningful latent components. Methods like Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) have been proposed for this purpose. However, they are limited in the number of components they can extract and also do not have a provision to control the expressiveness" of the extracted components. In this paper, we present a learning formulation to address these limitations by employing the notion of sparsity. We start with the PLSA framework and use an entropic prior in a maximum a posteriori formulation to enforce sparsity. We show that this allows the extraction of overcomplete sets of latent components which better characterize the data. We present experimental evidence of the utility of such representations."

****

Modelling motion primitives and their timing in biologically executed movements
Ben Williams, Marc Toussaint, Amos J. Storkey

Biological movement is built up of sub-blocks or motion primitives. Such primitives provide a compact representation of movement which is also desirable in robotic control applications. We analyse handwriting data to gain a better understanding of use of primitives and their timings in biological movements. Inference of the shape and the timing of primitives can be done using a factorial HMM based model, allowing the handwriting to be represented in primitive timing space. This representation provides a distribution of spikes corresponding to the primitive activations, which can also be modelled using HMM architectures. We show how the coupling of the low level primitive model, and the higher level timing model during inference can produce good reconstructions of handwriting, with shared primitives for all characters modelled. This coupled model also captures the variance profile of the dataset which is accounted for by spike timing jitter. The timing code provides a compact representation of the movement while generating a movement without an explicit timing model produces a scribbling style of output.

****

Subspace-Based Face Recognition in Analog VLSI
Gonzalo Carvajal, Waldo Valenzuela, Miguel Figueroa

We describe an analog-VLSI neural network for face recognition based on subspace methods. The system uses a dimensionality-reduction network whose coefficients can be either programmed or learned on-chip to per- form PCA, or programmed to perform LDA. A second network with user- programmed coefficients performs classification with Manhattan distances. The system uses on-chip compensation techniques to reduce the effects of device mismatch. Using the ORL database with 12x12-pixel images, our circuit achieves up to 85% classification performance (98% of an equivalent software implementation).

****

Efficient multiple hyperparameter learning for log-linear models
Chuan-sheng Foo, Chuong B., Andrew Ng

Using multiple regularization hyperparameters is an effective method for managing model complexity in problems where input features have varying amounts of noise. While algorithms for choosing multiple hyperparameters are often used in neural networks and support vector machines, they are not common in structured prediction tasks, such as sequence labeling or parsing. In this paper, we consider the problem of learning regularization hyperparameters for log-linear models, a class of probabilistic models for structured prediction tasks which includes conditional random fields (CRFs). Using an implicit differentiation trick, we derive an efficient gradient-based method for learning Gaussian regularization priors with multiple hyperparameters. In both simulations and the real-world task of computational RNA secondary structure prediction, we find that multiple hyperparameter learning provides a significant boost in accuracy compared to models learned using only a single regularization hyperparameter.

****

Discovering Weakly-Interacting Factors in a Complex Stochastic Process
Charlie Frogner, Avi Pfeffer

Dynamic Bayesian networks are structured representations of stochastic pro- cesses. Despite their structure, exact inference in DBNs is generally intractable. One approach to approximate inference involves grouping the variables in the process into smaller factors and keeping independent beliefs over these factors. In this paper we present several techniques for decomposing a dynamic Bayesian network automatically to enable factored inference. We examine a number of fea- tures of a DBN that capture different types of dependencies that will cause error in factored inference. An empirical comparison shows that the most useful of these is a heuristic that estimates the mutual information introduced between factors by one step of belief propagation. In addition to features computed over entire factors, for efficiency we explored scores computed over pairs of variables. We present search methods that use these features, pairwise and not, to find a factor- ization, and we compare their results on several datasets. Automatic factorization extends the applicability of factored inference to large, complex models that are undesirable to factor by hand. Moreover, tests on real DBNs show that automatic factorization can achieve significantly lower error in some cases.
****

Stability Bounds for Non-i.i.d. Processes
Mehryar Mohri, Afshin Rostamizadeh

The notion of algorithmic stability has been used effectively in the past to derive tight generalization bounds. A key advantage of these bounds is that they are de- signed for specific learning algorithms, exploiting their particular properties. But, as in much of learning theory, existing stability analyses and bounds apply only in the scenario where the samples are independently and identically distributed (i.i.d.). In many machine learning applications, however, this assumption does not hold. The observations received by the learning algorithm often have some inherent temporal dependence, which is clear in system diagnosis or time series prediction problems. This paper studies the scenario where the observations are drawn from a station- ary beta-mixing sequence, which implies a dependence between observations that weaken over time. It proves novel stability-based generalization bounds that hold even with this more general setting. These bounds strictly generalize the bounds given in the i.i.d. case. We also illustrate their application in the case of several general classes of learning algorithms, including Support Vector Regression and Kernel Ridge Regression.
****

Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks
Ben Carterette, Rosie Jones

We propose a model that leverages the millions of clicks received by web search engines, to predict document relevance. This allows the comparison of ranking functions when clicks are available but complete relevance judgments are not. After an initial training phase using a set of relevance judgments paired with click data, we show that our model can predict the relevance score of documents that have not been judged. These predictions can be used to evaluate the performance of a search engine, using our novel formalization of the confidence of the standard evaluation metric discounted cumulative gain (DCG), so comparisons can be made across time and datasets. This contrasts with previous methods which can provide only pair-wise relevance judgements between results shown for the same query. When no relevance judgments are available, we can identify the better of two ranked lists up to 82% of the time, and with only two relevance judgments for each query, we can identify the better ranking up to 94% of the time. While our experiments are on sponsored search results, which is the financial backbone of web search, our method is general enough to be applicable to algorithmic web search results as well. Furthermore, we give an algorithm to guide the selection of additional documents to judge to improve confidence.
****

Efficient Bayesian Inference for Dynamically Changing Graphs
Ozgur Sumer, Umut Acar, Alexander Ihler, Ramgopal Mettu

Motivated by stochastic systems in which observed evidence and conditional de- pendencies between states of the network change over time, and certain quantities

of interest (marginal distributions, likelihood estimates etc.) must be updated, we study the problem of adaptive inference in tree-structured Bayesian networks. We describe an algorithm for adaptive inference that handles a broad range of changes to the network and is able to maintain marginal distributions, MAP estimates, and data likelihoods in all expected logarithmic time. We give an implementation of our algorithm and provide experiments that show that the algorithm can yield up to two orders of magnitude speedups on answering queries and responding to dy- namic changes over the sum-product algorithm.
****
Markov Chain Monte Carlo with People
Adam Sanborn, Thomas Griffiths
Many formal models of cognition implicitly use subjective probability distributions to capture the assumptions of human learners. Most applications of these models determine these distributions indirectly. We propose a method for directly determining the assumptions of human learners by sampling from subjective probability distributions. Using a correspondence between a model of human choice and Markov chain Monte Carlo (MCMC), we describe a method for sampling from the distributions over objects that people associate with different categories. In our task, subjects choose whether to accept or reject a proposed change to an object. The task is constructed so that these decisions follow an MCMC acceptance rule, defining a Markov chain for which the stationary distribution is the category distribution. We test this procedure for both artificial categories acquired in the laboratory, and natural categories acquired from experience.
****
Estimating disparity with confidence from energy neurons
Eric Tsang, Bertram Shi
Binocular fusion takes place over a limited region smaller than one degree of visual angle (Panum's fusional area), which is on the order of the range of preferred disparities measured in populations of disparity-tuned neurons in the visual cortex. However, the actual range of binocular disparities encountered in natural scenes ranges over tens of degrees. This discrepancy suggests that there must be a mechanism for detecting whether the stimulus disparity is either inside or outside of the range of the preferred disparities in the population. Here, we present a statistical framework to derive feature in a population of V1 disparity neuron to determine the stimulus disparity within the preferred disparity range of the neural population. When optimized for natural images, it yields a feature that can be explained by the normalization which is a common model in V1 neurons. We further makes use of the feature to estimate the disparity in natural images. Our proposed model generates more correct estimates than coarse-to-fine multiple scales approaches and it can also identify regions with occlusion. The approach suggests another critical role for normalization in robust disparity estimation.
****
Locality and low-dimensions in the prediction of natural experience from fMRI
Francois Meyer, Greg Stephens
Functional Magnetic Resonance Imaging (fMRI) provides an unprecedented window into the complex functioning of the human brain, typically detailing the activity of thousands of voxels during hundreds of sequential time points. Unfortunately, the interpretation of fMRI is complicated due both to the relatively unknown connection between the hemodynamic response and neural activity and the unknown spatiotemporal characteristics of the cognitive patterns themselves. Here, we use data from the Experience Based Cognition competition to compare global and local methods of prediction applying both linear and nonlinear techniques of dimensionality reduction. We build global low dimensional representations of an fMRI dataset, using linear and nonlinear methods. We learn a set of time series that are implicit functions of the fMRI data, and predict the values of these times series in the future from the knowledge of the fMRI data only. We find effective, low-dimensional models based on the principal components of cognitive activity in classically-defined anatomical regions, the Brodmann Areas. Furthermore for some of the stimuli, the top predictive regions were stable across subjects and epis

odes, including WernickeÕs area for verbal instructions, visual cortex for facial and body features, and visual-temporal regions (Brodmann Area 7) for velocity. These interpretations and the relative simplicity of our approach provide a transparent and conceptual basis upon which to build more sophisticated techniques for fMRI decoding. To our knowledge, this is the first time that classical areas have been used in fMRI for an effective prediction of complex natural experience.
****
Configuration Estimates Improve Pedestrian Finding
Duan Tran, David Forsyth
Fair discriminative pedestrian finders are now available. In fact, these pedestrian finders make most errors on pedestrians in configurations that are uncommon in the training data, for example, mounting a bicycle. This is undesirable. However, the human configuration can itself be estimated discriminatively using structure learning. We demonstrate a pedestrian finder which first finds the most likely human pose in the window using a discriminative procedure trained with structure learning on a small dataset. We then present features (local histogram of oriented gradient and local PCA of gradient) based on that configuration to an SVM classifier. We show, using the INRIA Person dataset, that estimates of configuration significantly improve the accuracy of a discriminative pedestrian finder.
****
A General Boosting Method and its Application to Learning Ranking Functions for Web Search
Zhaohui Zheng, Hongyuan Zha, Tong Zhang, Olivier Chapelle, Keke Chen, Gordon Sun
We present a general boosting method extending functional gradient boosting to optimize complex loss functions that are encountered in many machine learning problems. Our approach is based on optimization of quadratic upper bounds of the loss functions which allows us to present a rigorous convergence analysis of the algorithm. More importantly, this general framework enables us to use a standard regression base learner such as decision trees for fitting any loss function. We illustrate an application of the proposed method in learning ranking functions for Web search by combining both preference data and labeled data for training. We present experimental results for Web search using data from a commercial search engine that show significant improvements of our proposed methods over some existing methods.
****
Fixing Max-Product: Convergent Message Passing Algorithms for MAP LP-Relaxations
Amir Globerson, Tommi Jaakkola
We present a novel message passing algorithm for approximating the MAP problem in graphical models. The algorithm is similar in structure to max-product but unlike max-product it always converges, and can be proven to find the exact MAP solution in various settings. The algorithm is derived via block coordinate descent in a dual of the LP relaxation of MAP, but does not require any tunable parameters such as step size or tree weights. We also describe a generalization of the method to cluster based potentials. The new method is tested on synthetic and real-world problems, and compares favorably with previous approaches.
****
GRIFT: A graphical model for inferring visual classification features from human data
Michael Ross, Andrew Cohen
This paper describes a new model for human visual classification that enables the recovery of image features that explain human subjects' performance on different visual classification tasks. Unlike previous methods, this algorithm does not model their performance with a single linear classifier operating on raw image pixels. Instead, it models classification as the combination of multiple feature detectors. This approach extracts more information about human visual classification than has been previously possible with other methods and provides a foundation for further exploration.
****

An in-silico Neural Model of Dynamic Routing through Neuronal Coherence

Devarajan Sridharan, Brian Percival, John Arthur, Kwabena A. Boahen

We describe a neurobiologically plausible model to implement dynamic routing using the concept of neuronal communication through neuronal coherence. The model has a three-tier architecture: a raw input tier, a routing control tier, and an invariant output tier. The correct mapping between input and output tiers is realized by an appropriate alignment of the phases of their respective background oscillations by the routing control units. We present an example architecture, im- plemented on a neuromorphic chip, that is able to achieve circular-shift invariance. A simple extension to our model can accomplish circular-shift dynamic routing with only O(N) connections, compared to O(N 2) connections required by tradi- tional models.

****

A general agnostic active learning algorithm

Sanjoy Dasgupta, Daniel J. Hsu, Claire Monteleoni

We present an agnostic active learning algorithm for any hypothesis class of bounded VC dimension under arbitrary data distributions. Most previ- ous work on active learning either makes strong distributional assumptions, or else is computationally prohibitive. Our algorithm extends the simple scheme of Cohn, Atlas, and Ladner [1] to the agnostic setting, using re- ductions to supervised learning that harness generalization bounds in a simple but subtle manner. We provide a fall-back guarantee that bounds the algorithm’s label complexity by the agnostic PAC sample complexity. Our analysis yields asymptotic label complexity improvements for certain hypothesis classes and distributions. We also demonstrate improvements experimentally.

****

Simplified Rules and Theoretical Analysis for Information Bottleneck Optimization and PCA with Spiking Neurons

Lars Buesing, Wolfgang Maass

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-authors prior to requesting a name change in the electronic proceedings.

****

Hidden Common Cause Relations in Relational Learning

Ricardo Silva, Wei Chu, Zoubin Ghahramani

When predicting class labels for objects within a relational database, it is often helpful to consider a model for relationships: this allows for information between class labels to be shared and to improve prediction performance. However, there are different ways by which objects can be related within a relational database. One traditional way corresponds to a Markov network structure: each existing relation is represented by an undirected edge. This encodes that, conditioned on input features, each object label is independent of other object labels given its neighbors in the graph. However, there is no reason why Markov networks should be the only representation of choice for symmetric dependence structures. Here we discuss the case when relationships are postulated to exist due to hidden com- mon causes. We discuss how the resulting graphical model differs from Markov networks, and how it describes different types of real-world relational processes. A Bayesian nonparametric classification model is built upon this graphical repre- sentation and evaluated with several empirical studies.

****

Modeling image patches with a directed hierarchy of Markov random fields

Simon Osindero, Geoffrey E. Hinton

We describe an efficient learning procedure for multilayer generative models that combine the best aspects of Markov random fields and deep, directed belief nets. The generative models can be learned one layer at a time and when learning is complete they have a very fast inference procedure for computing a good approximation to the posterior distribution in all of the hidden layers. Each hidden layer has its own MRF whose energy function is modulated by the top-down directed connections from the layer above. To generate from the model, each layer in turn must settle to equilibrium given its top-down input. We show that this type of model is good at capturing the statistics of patches of natural images.

****

The Generalized FITC Approximation

Andrew Naish-guzman, Sean Holden

We present an ef■cient generalization of the sparse pseudo-input Gaussian pro- cess (SPGP) model developed by Snelson and Ghahramani [1], applying it to binary classi■cation problems. By taking advantage of the SPGP prior covari- ance structure, we derive a numerically stable algorithm with O(N M 2) training complexity —asymptotically the same as related sparse methods such as the in- formative vector machine [2], but which more faithfully represents the posterior. We present experimental results for several benchmark problems showing that in many cases this allows an exceptional degree of sparsity without compromis- ing accuracy. Following [1], we locate pseudo-inputs by gradient ascent on the marginal likelihood, but exhibit occasions when this is likely to fail, for which we suggest alternative solutions.

****

Cooled and Relaxed Survey Propagation for MRFs

Hai Chieu, Wee Lee, Yee Teh

We describe a new algorithm, Relaxed Survey Propagation (RSP), for ■nding MAP con■gurations in Markov random ■elds. We compare its performance with state-of-the-art algorithms including the max-product belief propagation, its se- quential tree-reweighted variant, residual (sum-product) belief propagation, and tree-structured expectation propagation. We show that it outperforms all ap- proaches for Ising models with mixed couplings, as well as on a web person disambiguation task formulated as a supervised clustering problem.

****

A Spectral Regularization Framework for Multi-Task Structure Learning

Andreas Argyriou, Massimiliano Pontil, Yiming Ying, Charles Micchelli

Learning the common structure shared by a set of supervised tasks is an important practical and theoretical problem. Knowledge of this structure may lead to bet - ter generalization performance on the tasks and may also facilitate learning new tasks. We propose a framework for solving this problem, which is based on reg - ularization with spectral functions of matrices. This class of regularization prob- lems exhibits appealing computational properties and can be optimized ef(cid:2)ciently by an alternating minimization algorithm. In addition, we provide a necessary and suf(cid:2)cient condition for convexity of the regularizer. We analyze concrete ex- amples of the framework, which are equivalent to regularization with Lp matrix norms. Experiments on two real data sets indicate that the algorithm scales well with the number of tasks and improves on state of the art statistical performance.

****

CPR for CSPs: A Probabilistic Relaxation of Constraint Propagation

Luis E. Ortiz

This paper proposes constraint propagation relaxation (CPR), a probabilistic approach to classical constraint propagation that provides another view on the whole parametric family of survey propagation algorithms SP(ρ), ranging from belief propagation ($\rho = 0$) to (pure) survey propagation($\rho = 1$). More importantly, the approach elucidates the implicit, but fundamental assumptions underlying SP(ρ), thus shedding some light on its effectiveness and leading to applications beyond k-SAT.

****

Theoretical Analysis of Learning with Reward-Modulated Spike-Timing-Dependent Plasticity

Dejan Pecevski, Wolfgang Maass, Robert Legenstein

Reward-modulated spike-timing-dependent plasticity (STDP) has recently emerged as a candidate for a learning rule that could explain how local learning rules at single synapses support behaviorally relevant adaptive changes in com- plex networks of spiking neurons. However the potential and limitations of this learning rule could so far only be tested through computer simulations. This ar- ticle p rovides tools for an analytic treatment of reward-modulated STDP, which allow us to predict under which conditions reward-modulated STDP will be able to achieve a desired learning effect. In particular, we can produce in this way a theoretical explanation and a computer model for a fundamental experimental finding on biofeedback in monkeys (reported in [1]).

****

A New View of Automatic Relevance Determination

David Wipf, Srikantan Nagarajan

Automatic relevance determination (ARD), and the closely-related sparse Bayesian learning (SBL) framework, are effective tools for pruning large numbers of irrelevant features. However, popular update rules used for this process are either prohibitively slow in practice and/or heuristic in nature without proven convergence properties. This paper furnishes an alternative means of optimizing a general ARD cost function using an auxiliary function that can naturally be solved using a series of re-weighted L1 problems. The result is an efficient algorithm that can be implemented using standard convex programming toolboxes and is guaranteed to converge to a stationary point unlike existing methods. The analysis also leads to additional insights into the behavior of previous ARD updates as well as the ARD cost function. For example, the standard fixed-point updates of MacKay (1992) are shown to be iteratively solving a particular min-max problem, although they are not guaranteed to lead to a stationary point. The analysis also reveals that ARD is exactly equivalent to performing MAP estimation using a particular feature- and noise-dependent \textit{non-factorial} weight prior with several desirable properties over conventional priors with respect to feature selection. In particular, it provides a tighter approximation to the L0 quasi-norm sparsity measure than the L1 norm. Overall these results suggests alternative cost functions and update procedures for selecting features and promoting sparse solutions.

****

Sequential Hypothesis Testing under Stochastic Deadlines

Peter Frazier, Angela J. Yu

Most models of decision-making in neuroscience assume an infinite horizon, which yields an optimal solution that integrates evidence up to a fixed decision threshold; however, under most experimental as well as naturalistic behavioral settings, the decision has to be made before some finite deadline, which is often experienced as a stochastic quantity, either due to variable external constraints or internal timing uncertainty. In this work, we formulate this problem as sequential hypothesis testing under a stochastic horizon. We use dynamic programming tools to show that, for a large class of deadline distributions, the Bayes-optimal solution requires integrating evidence up to a threshold that declines monotonically over time. We use numerical simulations to illustrate the optimal policy in the special cases of a fixed deadline and one that is drawn from a gamma distribution.

****

Discriminative Log-Linear Grammars with Latent Variables

Slav Petrov, Dan Klein

We demonstrate that log-linear grammars with latent variables can be practically trained using discriminative methods. Central to efficient discriminative training is a hierarchical pruning procedure which allows feature expectations to be efficiently approximated in a gradient-based procedure. We compare L1 and L2 regularization and show that L1 regularization is superior, requiring fewer iterations to converge, and yielding sparser solutions. On full-scale treebank parsing experiments, the discriminative latent models outperform both the comparable generative latent models as well as the discriminative non-latent baselines.

****

HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation

Bing Zhao, Eric Xing

We present a novel paradigm for statistical machine translation (SMT), based on joint modeling of word alignment and the topical aspects underlying bilingual document pairs via a hidden Markov Bilingual Topic AdMixture (HM-BiTAM). In this n ew paradigm, parallel sentence-pairs from a parallel document-pair are coupled via a certain semantic-flow, to ensure coherence of topical context in the alignment of matching words between languages, during likelihood-based training of topic-dependent translational lexicons, as well as topic representations in each language. The resulting trained HM-BiTAM can not only display topic patterns like other methods such as LDA, but now for bilingual corpora; it also offers a principled way of inferring optimal translation in a context-dependent way. Our method integrates the conventional IBM Models based on HMM --- a key component for most of the state-of-the-art SMT systems, with the recently proposed BiTAM model, and we report an extensive empirical analysis (in many way complementary to the description-oriented of our method in three aspects: word alignment, bilingual topic representation, and translation.

****

Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs

Ambuj Tewari, Peter Bartlett

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-authors prior to requesting a name change in the electronic proceedings.

****

TrueSkill Through Time: Revisiting the History of Chess

Pierre Dangauthier, Ralf Herbrich, Tom Minka, Thore Graepel

We extend the Bayesian skill rating system TrueSkill to infer entire time series of skills of players by smoothing through time instead of filtering. The skill of each participating player, say, every year is represented by a latent skill variable which is affected by the relevant game outcomes that year, and coupled with the skill variables of the previous and subsequent year. Inference in the resulting factor graph is carried out by approximate message passing (EP) along the time series of skills. As before the system tracks the uncertainty about player skills, explicitly models draws, can deal with any number of competing entities and can infer individual skills from team results. We extend the system to estimate player-specific draw mar- gins. Based on these models we present an analysis of the skill curves of important players in the history of chess over the past 150 years. Results include plots of players' lifetime skill development as well as the ability to compare the skills of different players across time. Our results indicate that a) the overall playing strength has increased over the past 150 years, and b) that modelling a player's ability to force a draw provides significantly better predictive power.

****

Topmoumoute Online Natural Gradient Algorithm

Nicolas Roux, Pierre-antoine Manzagol, Yoshua Bengio

Guided by the goal of obtaining an optimization algorithm that is both fast and yielding good generalization, we study the descent direction maximizing the decrease in generalization error or the probability of not increasing generalization error. The surprising result is that from both the Bayesian and frequentist perspectives this can yield the natural gradient direction. Although that direction can be very expensive to compute we develop an efficient, general, online approximation to the natural gradient descent which is suited to large scale problems. We report experimental results showing much faster convergence in computation time and in number of iterations with TONGA (Topmoumoute Online natural Gradient Algorithm) than with stochastic gradient descent, even on very large datasets.

****

Learning the structure of manifolds using random projections

Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, Nakul Verma

We present a simple variant of the k-d tree which automatically adapts to intrinsic low dimensional structure in data.

****

Learning Monotonic Transformations for Classification

Andrew Howard, Tony Jebara

A discriminative method is proposed for learning monotonic transforma- tions of the training data while jointly estimating a large-margin classi(cid:12)er. In many domains such as document classi(cid:12)cation, image histogram classi(cid:12)- cation and gene microarray experiments, (cid:12)xed monotonic transformations can be useful as a preprocessing step. However, most classi(cid:12)ers only explore these transformations through manual trial and error or via prior domain knowledge. The proposed method learns monotonic transformations auto- matically while training a large-margin classi(cid:12)er without any prior knowl- edge of the domain. A monotonic piecewise linear function is learned which transforms data for subsequent processing by a linear hyperplane classi(cid:12)er. Two algorithmic implementations of the method are formalized. The (cid:12)rst solves a convergent alternating sequence of quadratic and linear programs until it obtains a locally optimal solution. An improved algorithm is then derived using a convex semide(cid:12)nite relaxation that overcomes initializa- tion issues in the greedy optimization problem. The e(cid:11)ectiveness of these learned transformations on synthetic problems, text data and image data is demonstrated.

****

Combined discriminative and generative articulated pose and non-rigid shape estimation

Leonid Sigal, Alexandru Balan, Michael Black

Estimation of three-dimensional articulated human pose and motion from images is a central problem in computer vision. Much of the previous work has been limited by the use of crude generative models of humans represented as articu- lated collections of simple parts such as cylinders. Automatic initialization of such models has proved dif■cult and most approaches assume that the size and shape of the body parts are known a priori. In this paper we propose a method for automatically recovering a detailed parametric model of non-rigid body shape and pose from monocular imagery. Speci■cally, we represent the body using a param- eterized triangulated mesh model that is learned from a database of human range scans. We demonstrate a discriminative method to directly recover the model pa- rameters from monocular images using a conditional mixture of kernel regressors. This predicted pose and shape are used to initialize a generative model for more detailed pose and shape estimation. The resulting approach allows fully automatic pose and shape recovery from monocular and multi-camera imagery. Experimen- tal results show that our method is capable of robustly recovering articulated pose, shape and biometric measurements (e.g. height, weight, etc.) in both calibrated and uncalibrated camera environments.

****

Multiple-Instance Active Learning

Burr Settles, Mark Craven, Soumya Ray

In a multiple instance (MI) learning problem, instances are naturally organized into bags and it is the bags, instead of individual instances, that are labeled for training. MI learners assume that every instance in a bag labeled negative is actually negative, whereas at least one instance in a bag labeled positive is actually positive. We present a framework for active learning in the multiple-instance setting. In particular, we consider the case in which an MI learner is allowed to selectively query unlabeled instances in positive bags. This approach is well motivated in domains in which it is inexpensive to acquire bag labels and possible, but expensive, to acquire instance labels. We describe a method for learning from labels at mixed levels of granularity, and introduce two active query selection strategies motivated by the MI setting. Our experiments show that learning from instance labels can significantly improve performance of a basic MI learning algorithm in two multiple-instance domains: content-based image recognition and text classification.

****

Semi-Supervised Multitask Learning

Qiuhua Liu, Xuejun Liao, Lawrence Carin

A semi-supervised multitask learning (MTL) framework is presented, in which M parameterized semi-supervised classi■ers, each associated with one of M par- tially labeled data manifolds, are learned jointly under the constraint of a soft- sharing prior imposed over the parameters of the classi■ers. The unlabeled data ar e utilized by basing classifier learning on neighborhoods, induced by a Markov random walk over a graph representation of each manifold. Experimental results on real data sets demonstrate that semi-supervised MTL yields significant im- provements in generalization performance over either semi-supervised single-task learning (STL) or supervised MTL.

****

Contraction Properties of VLSI Cooperative Competitive Neural Networks of Spiking Neurons

Emre Neftci, Elisabetta Chicca, Giacomo Indiveri, Jean-jeacques Slotine, Rodney Douglas

A non–linear dynamic system is called contracting if initial conditions are forgotten exponentially fast, so that all trajectories converge to a single trajectory. We use contraction theory to derive an upper bound for the strength of recurrent connections that guarantees contraction for complex neural networks. Specifi- cally, we apply this theory to a special class of recurrent networks, often called Cooperative Competitive Networks (CCNs), which are an abstract representation of the cooperative-competitive connectivity observed in cortex. This specific type of network is believed to play a major role in shaping cortical responses and se- lecting the relevant signal among distractors and noise. In this paper, we analyze contraction of combined CCNs of linear threshold units and verify the results of our analysis in a hybrid analog/digital VLSI CCN comprising spiking neurons and dynamic synapses.

****

Mining Internet-Scale Software Repositories

Erik Linstead, Paul Rigor, Sushil Bajracharya, Cristina Lopes, Pierre Baldi

Large repositories of source code create new challenges and opportunities for statistical machine learning. Here we first develop an infrastructure for the automated crawling, parsing, and database storage of open source software. The infrastructure allows us to gather Internet-scale source code. For instance, in one experiment, we gather 4,632 java projects from SourceForge and Apache totaling over 38 million lines of code from 9,250 developers. Simple statistical analyses of the data first reveal robust power-law behavior for package, SLOC, and method call distributions. We then develop and apply unsupervised author-topic, probabilistic models to automatically discover the topics embedded in the code and extract topic-word and author-topic distributions. In addition to serving as a convenient summary for program function and developer activities, these and other related distributions provide a statistical and information-theoretic basis for quantifying and analyzing developer similarity and competence, topic scattering, and document tangling, with direct applications to software engineering. Finally, by combining software textual content with structural information captured by our CodeRank approach, we are able to significantly improve software retrieval performance, increasing the AUC metric to 0.86-- roughly 10-30% better than previous approaches based on text alone.

****

Optimal models of sound localization by barn owls

Brian Fischer

Sound localization by barn owls is commonly modeled as a matching procedure where localization cues derived from auditory inputs are compared to stored templates. While the matching models can explain properties of neural responses, no model explains how the owl resolves spatial ambiguity in the localization cues to produce accurate localization near the center of gaze. Here, we examine two models for the barn owl's sound localization behavior. First, we consider a maximum likelihood estimator in order to further evaluate the cue matching model. Second, we consider a maximum a posteriori estimator to test if a Bayesian model with a prior that emphasizes directions near the center of gaze can reproduce the owl's localization behavior. We show that the maximum likelihood estimator can not reproduce the owl's behavior, while the maximum a posteriori estimator is able to match the behavior. This result suggests that the standard cue matching model will not be sufficient to explain sound localization behavior in the barn owl. The Bayesian model provides a new framework for analyzing sound localization in the

barn owl and leads to predictions about the owl's localization behavior.
****

Discriminative K-means for Clustering
Jieping Ye, Zheng Zhao, Mingrui Wu
We present a theoretical study on the discriminative clustering framework, recently proposed for simultaneous subspace selection via linear discriminant analysis (LDA) and clustering. Empirical results have shown its favorable performance in comparison with several other popular clustering algorithms. However, the inherent relationship between subspace selection and clustering in this framework is not well understood, due to the iterative nature of the algorithm. We show in this paper that this iterative subspace selection and clustering is equivalent to kernel K-means with a specific kernel Gram matrix. This provides significant and new insights into the nature of this subspace selection procedure. Based on this equivalence relationship, we propose the Discriminative K-means (DisKmeans) algorithm for simultaneous LDA subspace selection and clustering, as well as an automatic parameter estimation procedure. We also present the nonlinear extension of DisKmeans using kernels. We show that the learning of the kernel matrix over a convex set of pre-specified kernel matrices can be incorporated into the clustering formulation. The connection between DisKmeans and several other clustering algorithms is also analyzed. The presented theories and algorithms are evaluated through experiments on a collection of benchmark data sets.
****

Heterogeneous Component Analysis
Shigeyuki Oba, Motoaki Kawanabe, Klaus-Robert Müller, Shin Ishii
In bioinformatics it is often desirable to combine data from various measurement sources and thus structured feature vectors are to be analyzed that possess different intrinsic blocking characteristics (e.g., different patterns of missing values, obser- vation noise levels, effective intrinsic dimensionalities). We propose a new ma- chine learning tool, heterogeneous component analysis (HCA), for feature extrac- tion in order to better understand the factors that underlie such complex structured heterogeneous data. HCA is a linear block-wise sparse Bayesian PCA based not only on a probabilistic model with block-wise residual variance terms but also on a Bayesian treatment of a block-wise sparse factor-loading matrix. We study vari- ous algorithms that implement our HCA concept extracting sparse heterogeneous structure by obtaining common components for the blocks and specific compo- nents within each block. Simulations on toy and bioinformatics data underline the usefulness of the proposed structured matrix factorization concept.
****

An Analysis of Inference with the Universum
Olivier Chapelle, Alekh Agarwal, Fabian Sinz, Bernhard Schölkopf
We study a pattern classification algorithm which has recently been proposed by Vapnik and coworkers. It builds on a new inductive principle which assumes that in addition to positive and negative data, a third class of data is available, termed the Universum. We assay the behavior of the algorithm by establishing links with Fisher discriminant analysis and oriented PCA, as well as with an SVM in a pro- jected subspace (or, equivalently, with a data-dependent reduced kernel). We also provide experimental results.
****

Exponential Family Predictive Representations of State
David Wingate, Satinder Baveja
In order to represent state in controlled, partially observable, stochastic dynamical systems, some sort of sufficient statistic for history is necessary. Predictive repre- sentations of state (PSRs) capture state as statistics of the future. We introduce a new model of such systems called the “Exponential family PSR,” which defines as state the time-varying parameters of an exponential family distribution which models n sequential observations in the future. This choice of state representation explicitly connects PSRs to state-of-the-art probabilistic modeling, which allows us to take advantage of current efforts in high-dimensional density estimation, and in particular, graphical models and maximum entropy mode

ls. We present a pa- rameter learning algorithm based on maximum likelihood, and we show how a variety of current approximate inference methods apply. We evaluate the qual- ity of our model with reinforcement learning by directly evaluating the control performance of the model.
****
One-Pass Boosting
Zafer Barutcuoglu, Phil Long, Rocco Servedio
This paper studies boosting algorithms that make a single pass over a set of base classi(cid:2)ers. We (cid:2)rst analyze a one-pass algorithm in the setting of boosting with diverse base classi(cid:2)ers. Our guarantee is the same as the best proved for any boosting algo- rithm, but our one-pass algorithm is much faster than previous approaches. We next exhibit a random source of examples for which a (cid:147)picky(cid:148) variant of Ad- aBoost that skips poor base classi(cid:2)ers can outperform the standard AdaBoost al- gorithm, which uses every base classi(cid:2)er, by an exponential factor. Experiments with Reuters and synthetic data show that one-pass boosting can sub- stantially improve on the accuracy of Naive Bayes, and that picky boosting can sometimes lead to a further improvement in accuracy.
****
The Value of Labeled and Unlabeled Examples when the Model is Imperfect
Kaushik Sinha, Mikhail Belkin
Semi-supervised learning, i.e. learning from both labeled and unlabeled data has received signi(cid:2)cant attention in the machine learning literature in recent years. Still our understanding of the theoretical foundations of the usefulness of unla- beled data remains somewhat limited. The simplest and the best understood sit- uation is when the data is described by an identi(cid:2)able mixture model, and where each class comes from a pure component. This natural setup and its implications ware analyzed in [11, 5]. One important result was that in certain regimes, labeled data becomes exponentially more valuable than unlabeled data. However, in most realistic situations, one would not expect that the data comes from a parametric mixture distribution with identi(cid:2)able components. There have been recent efforts to analyze the non-parametric situation, for example, (cid:147)cluster(cid:148) and (cid:147)manifold(cid:148) assumptions have been suggested as a basis for analysis. Still, a satisfactory and fairly complete theoretical understanding of the nonparametric problem, similar to that in [11, 5] has not yet been developed. In this paper we investigate an intermediate situation, when the data comes from a probability distribution, which can be modeled, but not perfectly, by an identi(cid:2)able mixture distribution. This seems applicable to many situation, when, for example, a mixture of Gaussians is used to model the data. the contribution of this paper is an analysis of the role of labeled and unlabeled data depending on the amount of imperfection in the model.
****
On higher-order perceptron algorithms
Claudio Gentile, Fabio Vitale, Cristian Brotto
A new algorithm for on-line learning linear-threshold functions is proposed which efficiently combines second-order statistics about the data with the logarithmic behavior" of multiplicative/dual-norm algorithms. An initial theoretical analysis is provided suggesting that our algorithm might be viewed as a standard Perceptron algorithm operating on a transformed sequence of examples with improved margin properties. We also report on experiments carried out on datasets from diverse domains, with the goal of comparing to known Perceptron algorithms (first-order, second-order, additive, multiplicative). Our learning procedure seems to generalize quite well, and converges faster than the corresponding multiplicative baseline algorithms."
****
Adaptive Online Gradient Descent
Elad Hazan, Alexander Rakhlin, Peter Bartlett
We study the rates of growth of the regret in online convex optimization. First, we show that a simple extension of the algorithm of Hazan et al eliminates the need for a priori knowledge of the lower bound on the second derivatives of the

observed functions. We then provide an algorithm, Adaptive Online Gradient Descent, which interpolates between the results of Zinkevich for linear functions and of Hazan et al for strongly convex functions, achieving intermediate rates T and log T . Furthermore, we show strong optimality of the algorithm. between Finally, we provide an extension of our results to general norms.

****

Fast Variational Inference for Large-scale Internet Diagnosis

Emre Kiciman, David Maltz, John Platt

Web servers on the Internet need to maintain high reliability, but the cause of intermittent failures of web transactions is non-obvious. We use Bayesian inference to diagnose problems with web services. This diagnosis problem is far larger than any previously attempted: it requires inference of 10^4 possible faults from 10^5 observations. Further, such inference must be performed in less than a second. Inference can be done at this speed by combining a variational approximation, a mean-field approximation, and the use of stochastic gradient descent to optimize a variational cost function. We use this fast inference to diagnose a time series of anomalous HTTP requests taken from a real web service. The inference is fast enough to analyze network logs with billions of entries in a matter of hours.

****

Learning and using relational theories

Charles Kemp, Noah Goodman, Joshua Tenenbaum

Much of human knowledge is organized into sophisticated systems that are often called intuitive theories. We propose that intuitive theories are mentally represented in a logical language, and that the subjective complexity of a theory is determined by the length of its representation in this language. This complexity measure helps to explain how theories are learned from relational data, and how they support inductive inferences about unobserved relations. We describe two experiments that test our approach, and show that it provides a better account of human learning and reasoning than an approach developed by Goodman [1].

****

The Infinite Markov Model

Daichi Mochihashi, Eiichiro Sumita

We present a nonparametric Bayesian method of estimating variable order Markov processes up to a theoretically infinite order. By extending a stick-breaking prior, which is usually defined on a unit interval, “vertically” to the trees of infinite depth associated with a hierarchical Chinese restaurant process, our model directly infers the hidden orders of Markov dependencies from which each symbol originated. Experiments on character and word sequences in natural language showed that the model has a comparative performance with an exponentially large full-order model, while computationally much efficient in both time and space. We expect that this basic model will also extend to the variable order hierarchical clustering of general data.

****

Retrieved context and the discovery of semantic structure

Vinayak Rao, Marc Howard

Semantic memory refers to our knowledge of facts and relationships between concepts. A successful semantic memory depends on inferring relationships between items that are not explicitly taught. Recent mathematical modeling of episodic memory argues that episodic recall relies on retrieval of a gradually-changing rep - resentation of temporal context. We show that retrieved context enables the de - velopment of a global memory space that reflects relationships between all items that have been previously learned. When newly-learned information is integrated into this structure, it is placed in some relationship to all other items, even if that relationship has not been explicitly learned. We demonstrate this effect for global semantic structures shaped topologically as a ring, and as a two-dimensional sheet. We also examined the utility of this learning algorithm for learning a more realistic semantic space by training it on a large pool of synonym pairs. Retrieved context enabled the model to “infer” relationships between synonym pairs that had not yet been presented.

****
Active Preference Learning with Discrete Choice Data
Brochu Eric, Nando Freitas, Abhijeet Ghosh

We propose an active learning algorithm that learns a continuous valuation model from discrete preferences. The algorithm automatically decides what items are best presented to an individual in order to find the item that they value highly in as few trials as possible, and exploits quirks of human psychology to minimize time and cognitive burden. To do this, our algorithm maximizes the expected improvement at each query without accurately modelling the entire valuation surface, which would be needlessly expensive. The problem is particularly difficult because the space of choices is infinite. We demonstrate the effectiveness of the new algorithm compared to related active learning methods. We also embed the algorithm within a decision making tool for assisting digital artists in rendering materials. The tool finds the best parameters while minimizing the number of queries.

****
Bayesian binning beats approximate alternatives: estimating peri-stimulus time histograms
Dominik Endres, Mike Oram, Johannes Schindelin, Peter Foldiak

The peristimulus time historgram (PSTH) and its more continuous cousin, the spike density function (SDF) are staples in the analytic toolkit of neurophysiologists. The former is usually obtained by binning spiketrains, whereas the standard method for the latter is smoothing with a Gaussian kernel. Selection of a bin with or a kernel size is often done in an relatively arbitrary fashion, even though there have been recent attempts to remedy this situation \cite{ShimazakiBinningNIPS2006,ShimazakiBinningNECO2007}. We develop an exact Bayesian, generative model approach to estimating PSHTs and demonstate its superiority to competing methods. Further advantages of our scheme include automatic complexity control and error bars on its predictions.

****
Online Linear Regression and Its Application to Model-Based Reinforcement Learning
Alexander Strehl, Michael Littman

We provide a provably efficient algorithm for learning Markov Decision Processes (MDPs) with continuous state and action spaces in the online setting. Specifically, we take a model-based approach and show that a special type of online linear regression allows us to learn MDPs with (possibly kernalized) linearly parameterized dynamics. This result builds on Kearns and Singh's work that provides a provably efficient algorithm for finite state MDPs. Our approach is not restricted to the linear setting, and is applicable to other classes of continuous MDPs.

****
Transfer Learning using Kolmogorov Complexity: Basic Theory and Empirical Evaluations
M. Mahmud, Sylvian Ray

In transfer learning we aim to solve new problems using fewer examples using information gained from solving related problems. Transfer learning has been successful in practice, and extensive PAC analysis of these methods has been de- veloped. However it is not yet clear how to define relatedness between tasks. This is considered as a major problem as it is conceptually troubling and it makes it unclear how much information to transfer and when and how to transfer it. In this paper we propose to measure the amount of information one task contains about another using conditional Kolmogorov complexity between the tasks. We show how existing theory neatly solves the problem of measuring relatedness and transferring the ‘right’ amount of information in sequential transfer learning in a Bayesian setting. The theory also suggests that, in a very formal and precise sense, no other reasonable transfer method can do much better than our Kolmogorov Complexity theoretic transfer method, and that sequential transfer is always justi- fied. We also develop a practical approximation to the method and use it to transfer information between 8 arbitrarily chosen databases from the UCI ML repository.

****

McRank: Learning to Rank Using Multiple Classification and Gradient Boosting
Ping Li, Qiang Wu, Christopher Burges
We cast the ranking problem as (1) multiple classification (“Mc”) (2) multiple or- dinal classification, which lead to computationally tractable learning algorithms for relevance ranking in Web search. We consider the DCG criterion (discounted cumulative gain), a standard quality measure in information retrieval. Our approach is motivated by the fact that perfect classifications result in perfect DCG scores and the DCG errors are bounded by classification errors. We propose using the Expected Relevance to convert class probabilities into ranking scores. The class probabilities are learned using a gradient boosting tree algorithm. Evalua- tions on large-scale datasets show that our approach can improve LambdaRank [5] and the regressions-based ranker [6], in terms of the (normalized) DCG scores. An efficient implementation of the boosting tree algorithm is also presented.
****
A Randomized Algorithm for Large Scale Support Vector Learning
Krishnan Kumar, Chiru Bhattacharya, Ramesh Hariharan
We propose a randomized algorithm for large scale SVM learning which solves the problem by iterating over random subsets of the data. Crucial to the algorithm for scalability is the size of the subsets chosen. In the context of text classification we show that, by using ideas from random projections, a sample size of O(log n) can be used to obtain a solution which is close to the optimal with a high probability. Experiments done on synthetic and real life data sets demonstrate that the algorithm scales up SVM learners, without loss in accuracy.
****
Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation
Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, Motoaki Kawanabe
When training and test samples follow different input distributions (i.e., the situation called \emph{covariate shift}), the maximum likelihood estimator is known to lose its consistency. For regaining consistency, the log-likelihood terms need to be weighted according to the \emph{importance} (i.e., the ratio of test and training input densities). Thus, accurately estimating the importance is one of the key tasks in covariate shift adaptation. A naive approach is to first estimate training and test input densities and then estimate the importance by the ratio of the density estimates. However, since density estimation is a hard problem, this approach tends to perform poorly especially in high dimensional cases. In this paper, we propose a direct importance estimation method that does not require the input density estimates. Our method is equipped with a natural model selection procedure so tuning parameters such as the kernel width can be objectively optimized. This is an advantage over a recently developed method of direct importance estimation. Simulations illustrate the usefulness of our approach.
****
The Price of Bandit Information for Online Optimization
Varsha Dani, Sham M. Kakade, Thomas Hayes
In the online linear optimization problem, a learner must choose, in each round, a decision from a set $D \subset Rn$ in order to minimize an (unknown and chang- ing) linear cost function. We present sharp rates of convergence (with respect to additive regret) for both the full information setting (where the cost function is revealed at the end of each round) and the bandit setting (where only the scalar cost incurred is revealed). In particular, this paper is concerned with the price of bandit information, by which we mean the ratio of the best achievable regret √ in the bandit setting to that in the full-information setting. For the full informa- tion case, the upper bound on the regret is O∗(nT), where n is the ambient √ dimension and T is the time horizon. For the bandit case, we present an algorithm which achieves O∗(n3/2 T) regret — all previous (nontrivial) bounds here were O(poly(n)T 2/3) or worse. It is striking that the convergence rate for the bandit setting is only a factor of n worse than in the full information case — in stark √ contrast to the K-arm bandit setting, where the gap in the depende

nce on K is T log K). We also present lower bounds showing that exponential (this gap is at least n, which we conjecture to be the correct order. The bandit algorithm we present can be implemented efficiently in special cases of particular interest, such as path planning and Markov Decision Problems.
****

Iterative Non-linear Dimensionality Reduction with Manifold Sculpting
Michael Gashler, Dan Ventura, Tony Martinez
Many algorithms have been recently developed for reducing dimensionality by projecting data onto an intrinsic non-linear manifold. Unfortunately, existing algorithms often lose significant precision in this transformation. Manifold Sculpting is a new algorithm that iteratively reduces dimensionality by simulating surface tension in local neighborhoods. We present several experiments that show Manifold Sculpting yields more accurate results than existing algorithms with both generated and natural data-sets. Manifold Sculpting is also able to benefit from both prior dimensionality reduction efforts.
****

Support Vector Machine Classification with Indefinite Kernels
Ronny Luss, Alexandre D'aspremont
In this paper, we propose a method for support vector machine classification using indefinite kernels. Instead of directly minimizing or stabilizing a nonconvex loss function, our method simultaneously finds the support vectors and a proxy kernel matrix used in computing the loss. This can be interpreted as a robust classification problem where the indefinite kernel matrix is treated as a noisy observation of the true positive semidefinite kernel. Our formulation keeps the problem convex and relatively large problems can be solved efficiently using the analytic center cutting plane method. We compare the performance of our technique with other methods on several data sets.
****

Learning with Transformation Invariant Kernels
Christian Walder, Olivier Chapelle
This paper considers kernels invariant to translation, rotation and dilation. We show that no non-trivial positive definite (p.d.) kernels exist which are radial and dilation invariant, only conditionally positive definite (c.p.d.) ones. Accordingly, we discuss the c.p.d. case and provide some novel analysis, including an elemen- tary derivation of a c.p.d. representer theorem. On the practical side, we give a support vector machine (s.v.m.) algorithm for arbitrary c.p.d. kernels. For the thin- plate kernel this leads to a classifier with only one parameter (the amount of regu- larisation), which we demonstrate to be as effective as an s.v.m. with the Gaussian kernel, even though the Gaussian involves a second parameter (the length scale).
****

A probabilistic model for generating realistic lip movements from speech
Gwenn Englebienne, Tim Cootes, Magnus Rattray
The present work aims to model the correspondence between facial motion and speech. The face and sound are modelled separately, with phonemes being the link between both. We propose a sequential model and evaluate its suitability for the generation of the facial animation from a sequence of phonemes, which we obtain from speech. We evaluate the results both by computing the error between generated sequences and real video, as well as with a rigorous double-blind test with human subjects. Experiments show that our model compares favourably to other existing methods and that the sequences generated are comparable to real video sequences.
****

Automatic Generation of Social Tags for Music Recommendation
Douglas Eck, Paul Lamere, Thierry Bertin-mahieux, Stephen Green
Social tags are user-generated keywords associated with some resource on the Web. In the case of music, social tags have become an important component of Web2.0" recommender systems, allowing users to generate playlists based on use-dependent terms such as "chill" or "jogging" that have been applied to particular songs. In this paper, we propose a method for predicting these social tags directly f

rom MP3 files. Using a set of boosted classifiers, we map audio features onto social tags collected from the Web. The resulting automatic tags (or "autotags") furnish information about music that is otherwise untagged or poorly tagged, allowing for insertion of previously unheard music into a social recommender. This avoids the ''cold-start problem'' common in such systems. Autotags can also be used to smooth the tag space from which similarities and recommendations are made by providing a set of comparable baseline tags for all tracks in a recommender system."

****

Learning to classify complex patterns using a VLSI network of spiking neurons

Srinjoy Mitra, Giacomo Indiveri, Stefano Fusi

We propose a compact, low power VLSI network of spiking neurons which can learn to classify complex patterns of mean firing rates on–line and in real–time. The network of integrate-and-fire neurons is connected by bistable synapses that can change their weight using a local spike–based plasticity mechanism. Learning is supervised by a teacher which provides an extra input to the output neurons during training. The synaptic weights are updated only if the current generated by the plastic synapses does not match the output desired by the teacher (as in the perceptron learning rule). We present experimental results that demonstrate how this VLSI network is able to robustly classify uncorrelated linearly separable spatial patterns of mean firing rates.

****

Efficient Convex Relaxation for Transductive Support Vector Machine

Zenglin Xu, Rong Jin, Jianke Zhu, Irwin King, Michael Lyu

We consider the problem of Support Vector Machine transduction, which involves a combinatorial problem with exponential computational complexity in the number of unlabeled examples. Although several studies are devoted to Transductive SVM, they suffer either from the high computation complexity or from the solutions of local optimum. To address this problem, we propose solving Transductive SVM via a convex relaxation, which converts the NP-hard problem to a semi-definite programming. Compared with the other SDP relaxation for Transductive SVM, the proposed algorithm is computationally more efficient with the number of free parameters reduced from O(n2) to O(n) where n is the number of examples. Empirical study with several benchmark data sets shows the promising performance of the proposed algorithm in comparison with other state-of-the-art implementations of Transductive SVM.

****

Message Passing for Max-weight Independent Set

Sujay Sanghavi, Devavrat Shah, Alan Willsky

We investigate the use of message-passing algorithms for the problem of finding the max-weight independent set (MWIS) in a graph. First, we study the perfor- mance of loopy max-product belief propagation. We show that, if it converges, the quality of the estimate is closely related to the tightness of an LP relaxation of the MWIS problem. We use this relationship to obtain sufficient conditions for correctness of the estimate. We then develop a modification of max-product – one that converges to an optimal solution of the dual of the MWIS problem. We also develop a simple iterative algorithm for estimating the max-weight independent set from this dual solution. We show that the MWIS estimate obtained using these two algorithms in conjunction is correct when the graph is bipartite and the MWIS is unique. Finally, we show that any problem of MAP estimation for probability distributions over finite domains can be reduced to an MWIS problem. We believe this reduction will yield new insights and algorithms for MAP estimation.

****

Boosting the Area under the ROC Curve

Phil Long, Rocco Servedio

We show that any weak ranker that can achieve an area under the ROC curve slightly better than 1/2 (which can be achieved by random guessing) can be effi- ciently boosted to achieve an area under the ROC curve arbitrarily close to 1. We further show that this boosting can be performed even in the presence of indepen- dent misclassification noise, given access to a noise-tolerant weak ranker.

****

Sparse Feature Learning for Deep Belief Networks

Marc'aurelio Ranzato, Y-lan Boureau, Yann Cun

Unsupervised learning algorithms aim to discover the structure hidden in the data, and to learn representations that are more suitable as input to a supervised machine than the raw input. Many unsupervised methods are based on reconstructing the input from the representation, while constraining the representation to have certain desirable properties (e.g. low dimension, sparsity, etc). Others are based on approximating density by stochastically reconstructing the input from the representation. We describe a novel and efficient algorithm to learn sparse representations, and compare it theoretically and experimentally with a similar machines trained probabilistically, namely a Restricted Boltzmann Machine. We propose a simple criterion to compare and select different unsupervised machines based on the trade-off between the reconstruction error and the information content of the representation. We demonstrate this method by extracting features from a dataset of handwritten numerals, and from a dataset of natural image patches. We show that by stacking multiple levels of such machines and by training sequentially, high-order dependencies between the input variables can be captured.

****

Receding Horizon Differential Dynamic Programming

Yuval Tassa, Tom Erez, William Smart

The control of high-dimensional, continuous, non-linear systems is a key problem in reinforcement learning and control. Local, trajectory-based methods, using techniques such as Differential Dynamic Programming (DDP) are not directly subject to the curse of dimensionality, but generate only local controllers. In this paper, we introduce Receding Horizon DDP (RH-DDP), an extension to the classic DDP algorithm, which allows us to construct stable and robust controllers based on a library of local-control trajectories. We demonstrate the effectiveness of our approach on a series of high-dimensional control problems using a simulated multi-link swimming robot. These experiments show that our approach effectively circumvents dimensionality issues, and is capable of dealing effectively with problems with (at least) 34 state and 14 action dimensions.

****

A Risk Minimization Principle for a Class of Parzen Estimators

Kristiaan Pelckmans, Johan Suykens, Bart Moor

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-authors prior to requesting a name change in the electronic proceedings.

****

Managing Power Consumption and Performance of Computing Systems Using Reinforcement Learning

Gerald Tesauro, Rajarshi Das, Hoi Chan, Jeffrey Kephart, David Levine, Freeman Rawson, Charles Lefurgy

Electrical power management in large-scale IT systems such as commercial data- centers is an application area of rapidly growing interest from both an economic and ecological perspective, with billions of dollars and millions of metric tons of CO2 emissions at stake annually. Businesses want to save power without sacrificing performance. This paper presents a reinforcement learning approach to simultaneous online management of both performance and power consumption. We apply RL in a realistic laboratory testbed using a Blade cluster and dynam- ically varying HTTP workload running on a commercial web applications mid- dleware platform. We embed a CPU frequency controller in the Blade servers’ firmware, and we train policies for this controller using a multi-criteria reward signal depending on both application performance and CPU power consumption. Our testbed scenario posed a number of challenges to successful use of RL, in- cluding multiple disparate reward functions, limited decision sampling rates, and pathologies arising when using multiple sensor readings as state variables. We describe innovative practical solutions to these challenges, and demonstrate clear performance improvements over both hand-designed policies as well as obvious “cookbook” RL impleme ntations.

****

A Game-Theoretic Approach to Apprenticeship Learning
Umar Syed, Robert E. Schapire

We study the problem of an apprentice learning to behave in an environment with an unknown reward function by observing the behavior of an expert. We follow on the work of Abbeel and Ng [1] who considered a framework in which the true reward function is assumed to be a linear combination of a set of known and observable features. We give a new algorithm that, like theirs, is guaranteed to learn a policy that is nearly as good as the expert's, given enough examples. However, unlike their algorithm, we show that ours may produce a policy that is substantially better than the expert's. Moreover, our algorithm is computationally faster, is easier to implement, and can be applied even in the absence of an expert. The method is based on a game-theoretic view of the problem, which leads naturally to a direct application of the multiplicative-weights algorithm of Freund and Schapire [2] for playing repeated matrix games. In addition to our formal presentation and analysis of the new algorithm, we sketch how the method can be applied when the transition function itself is unknown, and we provide an experimental demonstration of the algorithm on a toy video-game environment.

****

Scene Segmentation with CRFs Learned from Partially Labeled Images
Bill Triggs, Jakob Verbeek

Conditional Random Fields (CRFs) are an effective tool for a variety of different data segmentation and labeling tasks including visual scene interpretation, which seeks to partition images into their constituent semantic-level regions and assign appropriate class labels to each region. For accurate labeling it is important to capture the global context of the image as well as local information. We in- troduce a CRF based scene labeling model that incorporates both local features and features aggregated over the whole image or large sections of it. Secondly, traditional CRF learning requires fully labeled datasets which can be costly and troublesome to produce. We introduce a method for learning CRFs from datasets with many unlabeled nodes by marginalizing out the unknown labels so that the log-likelihood of the known ones can be maximized by gradient ascent. Loopy Belief Propagation is used to approximate the marginals needed for the gradi- ent and log-likelihood calculations and the Bethe free-energy approximation to the log-likelihood is monitored to control the step size. Our experimental results show that effective models can be learned from fragmentary labelings and that incorporating top-down aggregate features significantly improves the segmenta- tions. The resulting segmentations are compared to the state-of-the-art on three different image datasets.

****

Measuring Neural Synchrony by Message Passing
Justin Dauwels, François Vialatte, Tomasz Rutkowski, Andrzej Cichocki

A novel approach to measure the interdependence of two time series is proposed, referred to as “stochastic event synchrony” (SES); it quantifies the alignment of two point processes by means of the following parameters: time delay, variance of the timing jitter, fraction of “spurious” events, and average similarity of events. SES may be applied to generic one-dimensional and multi-dimensional point pro- cesses, however, the paper mainly focusses on point processes in time-frequency domain. The average event similarity is in that case described by two parameters: the average frequency offset between events in the time-frequency plane, and the variance of the frequency offset (“frequency jitter”); SES then consists of five pa- rameters in total. Those parameters quantify the synchrony of oscillatory events, and hence, they provide an alternative to existing synchrony measures that quan- tify amplitude or phase synchrony. The pairwise alignment of point processes is cast as a statistical inference problem, which is solved by applying the max- product algorithm on a graphical model. The SES parameters are determined from the resulting pairwise alignment by maximum a posteriori (MAP) estimation. The proposed interdependence measure is applied to the problem of detecting anoma- lies in EEG synchrony of Mild Cognitive Impairment (MCI) patients; th e results indicate that SES significantly improves the sensitivity of EEG in detecting MCI.

****

Discriminative Batch Mode Active Learning
Yuhong Guo, Dale Schuurmans

Active learning sequentially selects unlabeled instances to label with the goal of reducing the effort needed to learn a good classifier. Most previous studies in active learning have focused on selecting one unlabeled instance at one time while retraining in each iteration. However, single instance selection systems are unable to exploit a parallelized labeler when one is available. Recently a few batch mode active learning approaches have been proposed that select a set of most informative unlabeled instances in each iteration, guided by some heuristic scores. In this paper, we propose a discriminative batch mode active learning approach that formulates the instance selection task as a continuous optimization problem over auxiliary instance selection variables. The optimization is formuated to maximize the discriminative classification performance of the target classifier, while also taking the unlabeled data into account. Although the objective is not convex, we can manipulate a quasi-Newton method to obtain a good local solution. Our empirical studies on UCI datasets show that the proposed active learning is more effective than current state-of-the art batch mode active learning algorithms.

****

Comparing Bayesian models for multisensory cue combination without mandatory integration
Ulrik Beierholm, Ladan Shams, Wei J., Konrad Koerding

Bayesian models of multisensory perception traditionally address the problem of estimating an underlying variable that is assumed to be the cause of the two sen- sory signals. The brain, however, has to solve a more general problem: it also has to establish which signals come from the same source and should be integrated, and which ones do not and should be segregated. In the last couple of years, a few models have been proposed to solve this problem in a Bayesian fashion. One of these has the strength that it formalizes the causal structure of sensory signals. We first compare these models on a formal level. Furthermore, we conduct a psy- chophysics experiment to test human performance in an auditory-visual spatial localization task in which integration is not mandatory. We find that the causal Bayesian inference model accounts for the data better than other models. Keywords: causal inference, Bayesian methods, visual perception.

****

What makes some POMDP problems easy to approximate?
Wee Lee, Nan Rong, David Hsu

Point-based algorithms have been surprisingly successful in computing approx- imately optimal solutions for partially observable Markov decision processes (POMDPs) in high dimensional belief spaces. In this work, we seek to understand the belief-space properties that allow some POMDP problems to be approximated efficiently and thus help to explain the point-based algorithms' success often ob- served in the experiments. We show that an approximately optimal POMDP so- lution can be computed in time polynomial in the covering number of a reachable belief space, which is the subset of the belief space reachable from a given belief point. We also show that under the weaker condition of having a small covering number for an optimal reachable space, which is the subset of the belief space reachable under an optimal policy, computing an approximately optimal solution is NP-hard. However, given a suitable set of points that "cover" an optimal reach- able space well, an approximate solution can be computed in polynomial time. The covering number highlights several interesting properties that reduce the com- plexity of POMDP planning in practice, e.g., fully observed state variables, beliefs with sparse support, smooth beliefs, and circulant state-transition matrices.

****

Boosting Algorithms for Maximizing the Soft Margin
Gunnar Rätsch, Manfred K. K. Warmuth, Karen Glocer
Gunnar R¨atsch

****

Gaussian Process Models for Link Analysis and Transfer Learning

Kai Yu, Wei Chu

In this paper we develop a Gaussian process (GP) framework to model a collection of reciprocal random variables defined on the \emph{edges} of a network. We show how to construct GP priors, i.e.,~covariance functions, on the edges of directed, undirected, and bipartite graphs. The model suggests an intimate connection between \emph{link prediction} and \emph{transfer learning}, which were traditionally considered two separate research topics. Though a straightforward GP inference has a very high complexity, we develop an efficient learning algorithm that can handle a large number of observations. The experimental results on several real-world data sets verify superior learning capacity.

****

Near-Maximum Entropy Models for Binary Neural Representations of Natural Images

Matthias Bethge, Philipp Berens

Maximum entropy analysis of binary variables provides an elegant way for study- ing the role of pairwise correlations in neural populations. Unfortunately, these approaches suffer from their poor scalability to high dimensions. In sensory cod- ing, however, high-dimensional data is ubiquitous. Here, we introduce a new approach using a near-maximum entropy model, that makes this type of analy- sis feasible for very high-dimensional data—the model parameters can be derived in closed form and sampling is easy. Therefore, our NearMaxEnt approach can serve as a tool for testing predictions from a pairwise maximum entropy model not only for low-dimensional marginals, but also for high dimensional measurements of more than thousand units. We demonstrate its usefulness by studying natural images with dichotomized pixel intensities. Our results indicate that the statistics of such higher-dimensional measurements exhibit additional structure that are not predicted by pairwise correlations, despite the fact that pairwise correlations ex- plain the lower-dimensional marginal statistics surprisingly well up to the limit of dimensionality where estimation of the full joint distribution is feasible.

****

Privacy-Preserving Belief Propagation and Sampling

Michael Kearns, Jinsong Tan, Jennifer Wortman

We provide provably privacy-preserving versions of belief propagation, Gibbs sampling, and other local algorithms – distributed multiparty protocols in which each party or vertex learns only its final local value, and absolutely nothing else.

****

Bayesian Co-Training

Shipeng Yu, Balaji Krishnapuram, Harald Steck, R. Rao, Rómer Rosales

We propose a Bayesian undirected graphical model for co-training, or more generally for semi-supervised multi-view learning. This makes explicit the previously unstated assumptions of a large class of co-training type algorithms, and also clarifies the circumstances under which these assumptions fail. Building upon new insights from this model, we propose an improved method for co-training, which is a novel co-training kernel for Gaussian process classifiers. The resulting approach is convex and avoids local-maxima problems, unlike some previous multi-view learning methods. Furthermore, it can automatically estimate how much each view should be trusted, and thus accommodate noisy or unreliable views. Experiments on toy data and real world data sets illustrate the benefits of this approach.

****

Supervised Topic Models

Jon Mcauliffe, David Blei

We introduce supervised latent Dirichlet allocation (sLDA), a statistical model of labelled documents. The model accommodates a variety of response types. We derive a maximum-likelihood procedure for parameter estimation, which relies on variational approximations to handle intractable posterior expectations. Prediction problems motivate this research: we use the fitted model to predict response values for new documents. We test sLDA on two real-world problems: movie ratings predicted from reviews, and web page popularity predicted from text descriptions. We illustrate the benefits of sLDA versus modern regularized regression, as well as versus an unsupervised LDA analysis followed by a separate regression.

****

A Kernel Statistical Test of Independence

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, Alex Smola

Although kernel measures of independence have been widely applied in machine learning (notably in kernel ICA), there is as yet no method to determine whether they have detected statistically significant dependence. We provide a novel test of the independence hypothesis for one particular kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). The resulting test costs O(m2), where m is the sample size. We demonstrate that this test outperforms established contingency table and functional correlation-based tests, and that this advantage is greater for multivariate data. Finally, we show the HSIC test also applies to text (and to structured data more generally), for which no other independence test presently exists.

****

Discriminative Keyword Selection Using Support Vector Machines

Fred Richardson, William Campbell

Many tasks in speech processing involve classification of long term characteristics of a speech segment such as language, speaker, dialect, or topic. A natural technique for determining these characteristics is to first convert the input speech into a sequence of tokens such as words, phones, etc. From these tokens, we can then look for distinctive phrases, keywords, that characterize the speech. In many applications, a set of distinctive keywords may not be known a priori. In this case, an automatic method of building up keywords from short context units such as phones is desirable. We propose a method for construction of keywords based upon Support Vector Machines. We cast the problem of keyword selection as a feature selection problem for n-grams of phones. We propose an alternating filter-wrapper method that builds successively longer keywords. Application of this method on a language recognition task shows that the technique produces interesting and significant qualitative and quantitative results.

****

Probabilistic Matrix Factorization

Andriy Mnih, Russ R. Salakhutdinov

Many existing approaches to collaborative filtering can neither handle very large datasets nor easily deal with users who have very few ratings. In this paper we present the Probabilistic Matrix Factorization (PMF) model which scales linearly with the number of observations and, more importantly, performs well on the large, sparse, and very imbalanced Netflix dataset. We further extend the PMF model to include an adaptive prior on the model parameters and show how the model capacity can be controlled automatically. Finally, we introduce a con- strained version of the PMF model that is based on the assumption that users who have rated similar sets of movies are likely to have similar preferences. The result- ing model is able to generalize considerably better for users with very few ratings. When the predictions of multiple PMF models are linearly combined with the predictions of Restricted Boltzmann Machines models, we achieve an error rate of 0.8861, that is nearly 7% better than the score of Netflix’s own system.

****

Density Estimation under Independent Similarly Distributed Sampling Assumptions

Tony Jebara, Yingbo Song, Kapil Thadani

A method is proposed for semiparametric estimation where parametric and non- parametric criteria are exploited in density estimation and unsupervised learning. This is accomplished by making sampling assumptions on a dataset that smoothly interpolate between the extreme of independently distributed (or id) sample data (as in nonparametric kernel density estimators) to the extreme of independent identically distributed (or iid) sample data. This article makes independent similarly distributed (or isd) sampling assumptions and interpolates between these two using a scalar parameter. The parameter controls a Bhattacharyya affinity pen alty between pairs of distributions on samples. Surprisingly, the isd method maintains certain consistency and unimodality properties akin to maximum likelihood estimation. The proposed isd scheme is an alternative for handling nonstationarity in data without making drastic hidden variable assumptions which often make estimation difficult and laden with local optima. Experiments in density estimation on a variety of datasets confirm the value of isd over iid estimation, id estimation and mixture modeling.

****

Efficient Inference for Distributions on Permutations

Jonathan Huang, Carlos Guestrin, Leonidas J. Guibas

Permutations are ubiquitous in many real world problems, such as voting, rankings and data association. Representing uncertainty over permutations is challenging, since there are n! possibilities, and typical compact representations such as graphical models cannot efficiently capture the mutual exclusivity constraints associated with permutations. In this paper, we use the “low-frequency” terms of a Fourier decomposition to represent such distributions compactly. We present Kronecker conditioning, a general and efficient approach for maintaining these distributions directly in the Fourier domain. Low order Fourier-based approximations can lead to functions that do not correspond to valid distributions. To address this problem, we present an efficient quadratic program defined directly in the Fourier domain to project the approximation onto a relaxed form of the marginal polytope. We demonstrate the effectiveness of our approach on a real camera-based multi-people tracking setting.

****

Fitted Q-iteration in continuous action-space MDPs

András Antos, Csaba Szepesvári, Rémi Munos

We consider continuous state, continuous action batch reinforcement learning where the goal is to learn a good policy from a sufficiently rich trajectory generated by another policy. We study a variant of fitted Q-iteration, where the greedy action selection is replaced by searching for a policy in a restricted set of candidate policies by maximizing the average action values. We provide a rigorous theoretical analysis of this algorithm, proving what we believe is the first finite-time bounds for value-function based algorithms for continuous state- and action-space problems.

****

Blind channel identification for speech dereverberation using l1-norm sparse learning

Yuanqing Lin, Jingdong Chen, Youngmoo Kim, Daniel Lee

Speech dereverberation remains an open problem after more than three decades of research. The most challenging step in speech dereverberation is blind channel identification (BCI). Although many BCI approaches have been developed, their performance is still far from satisfactory for practical applications. The main difficulty in BCI lies in finding an appropriate acoustic model, which not only can effectively resolve solution degeneracies due to the lack of knowledge of the source, but also robustly models real acoustic environments. This paper proposes a sparse acoustic room impulse response (RIR) model for BCI, that is, an acoustic RIR can be modeled by a sparse FIR filter. Under this model, we show how to formulate the BCI of a single-input multiple-output (SIMO) system into a l1-norm regularized least squares (LS) problem, which is convex and can be solved efficiently with guaranteed global convergence. The sparseness of solutions is controlled by l1-norm regularization parameters. We propose a sparse learning scheme that infers the optimal l1-norm regularization parameters directly from microphone observations under a Bayesian framework. Our results show that the proposed approach is effective and robust, and it yields source estimates in real acoustic environments with high fidelity to anechoic chamber measurements.

****

Predicting Brain States from fMRI Data: Incremental Functional Principal Component Regression

Sennay Ghebreab, Arnold Smeulders, Pieter Adriaans

We propose a method for reconstruction of human brain states directly from funct

ional neuroimaging data. The method extends the traditional multivariate regression analysis of discretized fMRI data to the domain of stochastic functional measurements, facilitating evaluation of brain responses to naturalistic stimuli and boosting the power of functional imaging. The method searches for sets of voxel timecourses that optimize a multivariate functional linear model in terms of Rsquare-statistic. Population based incremental learning is used to search for spatially distributed voxel clusters, taking into account the variation in Haemodynamic lag across brain areas and among subjects by voxel-wise non-linear registration of stimuli to fMRI data. The method captures spatially distributed brain responses to naturalistic stimuli without attempting to localize function. Application of the method for prediction of naturalistic stimuli from new and unknown fMRI data shows that the approach is capable of identifying distributed clusters of brain locations that are highly predictive of a specific stimuli.
****

Agreement-Based Learning
Percy S. Liang, Dan Klein, Michael Jordan
The learning of probabilistic models with many hidden variables and non- decomposable dependencies is an important and challenging problem. In contrast to traditional approaches based on approximate inference in a single intractable model, our approach is to train a set of tractable submodels by encouraging them to agree on the hidden variables. This allows us to capture non-decomposable aspects of the data while still maintaining tractability. We propose an objective function for our approach, derive EM-style algorithms for parameter estimation, and demonstrate their effectiveness on three challenging real-world learning tasks.
****

Extending position/phase-shift tuning to motion energy neurons improves velocity discrimination
Yiu Lam, Bertram Shi
We extend position and phase-shift tuning, concepts already well established in the disparity energy neuron literature, to motion energy neurons. We show that Reichardt-like detectors can be considered examples of position tuning, and that motion energy filters whose complex valued spatio-temporal receptive fields are space-time separable can be considered examples of phase tuning. By combining these two types of detectors, we obtain an architecture for constructing motion energy neurons whose center frequencies can be adjusted by both phase and posi- tion shifts. Similar to recently described neurons in the primary visual cortex, these new motion energy neurons exhibit tuning that is between purely space- time separable and purely speed tuned. We propose a functional role for this intermediate level of tuning by demonstrating that comparisons between pairs of these motion energy neurons can reliably discriminate between inputs whose velocities lie above or below a given reference velocity.
****

A Bayesian Framework for Cross-Situational Word-Learning
Noah Goodman, Joshua Tenenbaum, Michael Black
For infants, early word learning is a chicken-and-egg problem. One way to learn a word is to observe that it co-occurs with a particular referent across different situations. Another way is to use the social context of an utterance to infer the in- tended referent of a word. Here we present a Bayesian model of cross-situational word learning, and an extension of this model that also learns which social cues are relevant to determining reference. We test our model on a small corpus of mother-infant interaction and find it performs better than competing models. Fi- nally, we show that our model accounts for experimental phenomena including mutual exclusivity, fast-mapping, and generalization from social cues.
****

Parallelizing Support Vector Machines on Distributed Computers
Kaihua Zhu, Hao Wang, Hongjie Bai, Jian Li, Zhihuan Qiu, Hang Cui, Edward Chang
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-authors prior to requesting a name change in the electronic proceedings.

****

Object Recognition by Scene Alignment

Bryan Russell, Antonio Torralba, Ce Liu, Rob Fergus, William Freeman

Current object recognition systems can only recognize a limited number of object categories; scaling up to many categories is the next challenge. We seek to build a system to recognize and localize many different object categories in complex scenes. We achieve this through a simple approach: by matching the input im- age, in an appropriate representation, to images in a large training set of labeled images. Due to regularities in object identities across similar scenes, the retrieved matches provide hypotheses for object identities and locations. We build a prob- abilistic model to transfer the labels from the retrieval set to the input image. We demonstrate the effectiveness of this approach and study algorithm component contributions using held-out test sets from the LabelMe database.

****

A neural network implementing optimal state estimation based on dynamic spike train decoding

Omer Bobrowski, Ron Meir, Shy Shoham, Yonina Eldar

It is becoming increasingly evident that organisms acting in uncertain dynamical environments often employ exact or approximate Bayesian statistical calculations in order to continuously estimate the environmental state, integrate information from multiple sensory modalities, form predictions and choose actions. What is less clear is how these putative computations are implemented by cortical neural networks. An additional level of complexity is introduced because these networks observe the world through spike trains received from primary sensory afferents, rather than directly. A recent line of research has described mechanisms by which such computations can be implemented using a network of neurons whose activ- ity directly represents a probability distribution across the possible “world states”. Much of this work, however, uses various approximations, which severely re- strict the domain of applicability of these implementations. Here we make use of rigorous mathematical results from the theory of continuous time point process filtering, and show how optimal real-time state estimation and prediction may be implemented in a general setting using linear neural networks. We demonstrate the applicability of the approach with several examples, and relate the required network properties to the statistical nature of the environment, thereby quantify- ing the compatibility of a given network with its environment.

****

Computational Equivalence of Fixed Points and No Regret Algorithms, and Convergence to Equilibria

Elad Hazan, Satyen Kale

We study the relation between notions of game-theoretic equilibria which are based on stability under a set of deviations, and empirical equilibria which are reached by rational players. Rational players are modelled by players using no regret algorithms, which guarantee that their payoff in the long run is almost as much as the most they could hope to achieve by consistently deviating from the algorithm's suggested action. We show that for a given set of deviations over the strategy set of a player, it is possible to efficiently approximate fixed points of a given deviation if and only if there exist efficient no regret algorithms resistant to the deviations. Further, we show that if all players use a no regret algorithm, then the empirical distribution of their plays converges to an equilibrium.

****

Catching Change-points with Lasso

Céline Levy-leduc, Zaïd Harchaoui

We propose a new approach for dealing with the estimation of the location of change-points in one-dimensional piecewise constant signals observed in white noise. Our approach consists in reframing this task in a variable selection context. We use a penalized least-squares criterion with a l1-type penalty for this purpose. We prove that, in an appropriate asymptotic framework, this method provides consistent estimators of the change-points. Then, we explain how to implement this method in practice by combining the LAR algorithm and a reduced version of th e dynamic programming algorithm and we apply it to synthetic and real data.

****

Feature Selection Methods for Improving Protein Structure Prediction with Rosetta

Ben Blum, David Baker, Michael Jordan, Philip Bradley, Rhiju Das, David E. Kim

Rosetta is one of the leading algorithms for protein structure prediction today. It is a Monte Carlo energy minimization method requiring many random restarts to find structures with low energy. In this paper we present a resampling technique for structure prediction of small alpha/beta proteins using Rosetta. From an ini- tial round of Rosetta sampling, we learn properties of the energy landscape that guide a subsequent round of sampling toward lower-energy structures. Rather than attempt to fit the full energy landscape, we use feature selection methods—both L1-regularized linear regression and decision trees—to identify structural features that give rise to low energy. We then enrich these structural features in the second sampling round. Results are presented across a benchmark set of nine small al- pha/beta proteins demonstrating that our methods seldom impair, and frequently improve, Rosetta’s performance.

****

Selecting Observations against Adversarial Objectives

Andreas Krause, Brendan Mcmahan, Carlos Guestrin, Anupam Gupta

In many applications, one has to actively select among a set of expensive observa- tions before making an informed decision. Often, we want to select observations which perform well when evaluated with an objective function chosen by an adver- sary. Examples include minimizing the maximum posterior variance in Gaussian Process regression, robust experimental design, and sensor placement for outbreak detection. In this paper, we present the Submodular Saturation algorithm, a sim- ple and efficient algorithm with strong theoretical approximation guarantees for the case where the possible objective functions exhibit submodularity, an intuitive diminishing returns property. Moreover, we prove that better approximation al- gorithms do not exist unless NP-complete problems admit efficient algorithms. We evaluate our algorithm on several real-world problems. For Gaussian Process regression, our algorithm compares favorably with state-of-the-art heuristics de- scribed in the geostatistics literature, while being simpler, faster and providing theoretical guarantees. For robust experimental design, our algorithm performs favorably compared to SDP-based algorithms.

****

Spatial Latent Dirichlet Allocation

Xiaogang Wang, Eric Grimson

In recent years, the language model Latent Dirichlet Allocation (LDA), which clusters co-occurring words into topics, has been widely appled in the computer vision field. However, many of these applications have difficulty with modeling the spatial and temporal structure among visual words, since LDA assumes that a document is a bag-of-words''. It is also critical to properly designwords'' and “documents” when using a language model to solve vision problems. In this paper, we propose a topic model Spatial Latent Dirichlet Allocation (SLDA), which better encodes spatial structure among visual words that are essential for solving many vision problems. The spatial information is not encoded in the value of visual words but in the design of documents. Instead of knowing the partition of words into documents \textit{a priori}, the word-document assignment becomes a random hidden variable in SLDA. There is a generative procedure, where knowledge of spatial structure can be flexibly added as a prior, grouping visual words which are close in space into the same document. We use SLDA to discover objects from a collection of images, and show it achieves better performance than LDA.

****

Learning Visual Attributes

Vittorio Ferrari, Andrew Zisserman

We present a probabilistic generative model of visual attributes, together with an efficient learning algorithm. Attributes are visual qualities of objects, such as ‘red’, ‘striped’, or ‘spotted’. The model sees attributes as patterns of image segments, repeatedly sharing some characteristic properties. These can be any combination of appearance, shape, or the layout of segments within the pattern. Moreover, attributes with general appearance are taken into account, such as the pattern of alternation of any two colors which is characteristic for stripes. To enable learning from unsegmented training images, the model is learnt discriminatively, by optimizing a likelihood ratio. As demonstrated in the experimental evaluation, our model can learn in a weakly supervised setting and encompasses a broad range of attributes. We show that attributes can be learnt starting from a text query to Google image search, and can then be used to recognize the attribute and determine its spatial extent in novel real-world images.

****

Collapsed Variational Inference for HDP
Yee Teh, Kenichi Kurihara, Max Welling
A wide variety of Dirichlet-multinomial ‘topic’ models have found interesting applications in recent years. While Gibbs sampling remains an important method of inference in such models, variational techniques have certain advantages such as easy assessment of convergence, easy optimization without the need to maintain detailed balance, a bound on the marginal likelihood, and side-stepping of issues with topic-identifiability. The most accurate variational technique thus far, namely collapsed variational latent Dirichlet allocation, did not deal with model selection nor did it include inference for hyperparameters. We address both issues by gen- eralizing the technique, obtaining the first variational algorithm to deal with the hierarchical Dirichlet process and to deal with hyperparameters of Dirichlet vari- ables. Experiments show a significant improvement in accuracy.

****

Progressive mixture rules are deviation suboptimal
Jean-yves Audibert
We consider the learning task consisting in predicting as well as the best function in a finite reference set G up to the smallest possible additive term. If R(g) denotes the generalization error of a prediction function g, under reasonable assumptions on the loss function (typically satisfied by the least square loss when the output is bounded), it is known that the progressive mixture rule gn satisfies E R(gn) < min_{g in G} R(g) + Cst (log|G|)/n where n denotes the size of the training set, E denotes the expectation wrt the training set distribution. This work shows that, surprisingly, for appropriate reference sets G, the deviation convergence rate of the progressive mixture rule is only no better than Cst / sqrt{n}, and not the expected Cst / n. It also provides an algorithm which does not suffer from this drawback.

****

Experience-Guided Search: A Theory of Attentional Control
David Baldwin, Michael C. Mozer
People perform a remarkable range of tasks that require search of the visual environment for a target item among distractors. The Guided Search model (Wolfe, 1994, 2007), or GS, is perhaps the best developed psychological account of hu- man visual search. To prioritize search, GS assigns saliency to locations in the visual field. Saliency is a linear combination of activations from retinotopic maps representing primitive visual features. GS includes heuristics for setting the gain coefficient associated with each map. Variants of GS have formalized the notion of optimization as a principle of attentional control (e.g., Baldwin & Mozer, 2006; Cave, 1999; Navalpakkam & Itti, 2006; Rao et al., 2002), but every GS-like model must be ’dumbed down’ to match human data, e.g., by corrupting the saliency map with noise and by imposing arbitrary restrictions on gain modulation. We propose a principled probabilistic formulation of GS, called Experience-Guided Search (EGS), based on a generative model of the environment that makes three claims: (1) Feature detectors produce Poisson spike trains whose rates are conditioned on feature type and whether the feature belongs to a target or distractor; (2) the en- vironment and/or task is nonstationary and can change over a sequence of trials; and (3) a prior specifies that features are more likely to be present for target than for distractors. Through experience, EGS infers latent environment variables that determine the gains for guiding search. Control is thus c

ast as probabilistic infer- ence, not optimization. We show that EGS can replicate a range of human data from visual search, including data that GS does not address.
****
Hierarchical Penalization
Marie Szafranski, Yves Grandvalet, Pierre Morizet-mahoudeaux
Hierarchical penalization is a generic framework for incorporating prior informa- tion in the fitting of statistical models, when the explicative variables are organized in a hierarchical structure. The penalizer is a convex functional that performs soft selection at the group level, and shrinks variables within each group. This favors solutions with few leading terms in the final combination. The framework, orig- inally derived for taking prior knowledge into account, is shown to be useful in linear regression, when several parameters are used to model the influence of one feature, or in kernel regression, for learning multiple kernels. Keywords – Optimization: constrained and convex optimization. Supervised learning: regression, kernel methods, sparsity and feature selection.
****
Linear programming analysis of loopy belief propagation for weighted matching
Sujay Sanghavi, Dmitry Malioutov, Alan Willsky
Loopy belief propagation has been employed in a wide variety of applications with great empirical success, but it comes with few theoretical guarantees. In this paper we investigate the use of the max-product form of belief propagation for weighted matching problems on general graphs. We show that max-product converges to the correct answer if the linear programming (LP) relaxation of the weighted matching problem is tight and does not converge if the LP relaxation is loose. This provides an exact characterization of max-product performance and reveals connections to the widely used optimization technique of LP relaxation. In addition, we demonstrate that max-product is effective in solving practical weighted matching problems in a distributed fashion by applying it to the problem of self-organization in sensor networks.
****
Learning Horizontal Connections in a Sparse Coding Model of Natural Images
Pierre Garrigues, Bruno Olshausen
It has been shown that adapting a dictionary of basis functions to the statistics of natural images so as to maximize sparsity in the coefficients results in a set of dictionary elements whose spatial properties resemble those of V1 (primary visual cortex) receptive fields. However, the resulting sparse coefficients still exhibit pronounced statistical dependencies, thus violating the independence assumption of the sparse coding model. Here, we propose a model that attempts to capture the dependencies among the basis function coefficients by including a pairwise coupling term in the prior over the coefficient activity states. When adapted to the statistics of natural images, the coupling terms learn a combination of facilitatory and inhibitory interactions among neighboring basis functions. These learned interactions may offer an explanation for the function of horizontal connections in V1, and we discuss the implications of our findings for physiological experiments.
****
COFI RANK - Maximum Margin Matrix Factorization for Collaborative Ranking
Markus Weimer, Alexandros Karatzoglou, Quoc Le, Alex Smola
In this paper, we consider collaborative filtering as a ranking problem. We present a method which uses Maximum Margin Matrix Factorization and optimizes rank- ing instead of rating. We employ structured output prediction to optimize directly for ranking scores. Experimental results show that our method gives very good ranking scores and scales well on collaborative filtering tasks.
****
Infinite State Bayes-Nets for Structured Domains
Max Welling, Ian Porteous, Evgeniy Bart
A general modeling framework is proposed that unifies nonparametric-Bayesian models, topic-models and Bayesian networks. This class of infinite state Bayes nets (ISBN) can be viewed as directed networks of ‘hierarchical Dirichlet processes’ (

HDPs) where the domain of the variables can be structured (e.g. words in documents or features in images). We show that collapsed Gibbs sampling can be done efficiently in these models by leveraging the structure of the Bayes net and using the forward-filtering-backward-sampling algorithm for junction trees. Existing models, such as nested-DP, Pachinko allocation, mixed membership sto- chastic block models as well as a number of new models are described as ISBNs. Two experiments have been performed to illustrate these ideas.

****

Regularized Boost for Semi-Supervised Learning

Ke Chen, Shihai Wang

Semi-supervised inductive learning concerns how to learn a decision rule from a data set containing both labeled and unlabeled data. Several boosting algorithms have been extended to semi-supervised learning with various strategies. To our knowledge, however, none of them takes local smoothness constraints among data into account during ensemble learning. In this paper, we introduce a local smoothness regularizer to semi-supervised boosting algorithms based on the universal optimization framework of margin cost functionals. Our regularizer is applicable to existing semi-supervised boosting algorithms to improve their generalization and speed up their training. Comparative results on synthetic, benchmark and real world tasks demonstrate the effectiveness of our local smoothness regularizer. We discuss relevant issues and relate our regularizer to previous work.

****

Consistent Minimization of Clustering Objective Functions

Ulrike Luxburg, Stefanie Jegelka, Michael Kaufmann, Sébastien Bubeck

Clustering is often formulated as a discrete optimization problem. The objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the under- lying space. We argue that the discrete optimization approach usually does not achieve this goal. As an alternative, we suggest the paradigm of “nearest neighbor clustering”. Instead of selecting the best out of all partitions of the sample, it only considers partitions in some restricted function class. Using tools from statistical learning theory we prove that nearest neighbor clustering is statistically consis- tent. Moreover, its worst case complexity is polynomial by construction, and it can be implemented with small average case complexity using branch and bound.

****

The rat as particle filter

Aaron C. Courville, Nathaniel Daw

Although theorists have interpreted classical conditioning as a laboratory model of Bayesian belief updating, a recent reanalysis showed that the key features that theoretical models capture about learning are artifacts of averaging over subjects. Rather than learning smoothly to asymptote (reflecting, according to Bayesian models, the gradual tradeoff from prior to posterior as data accumulate), subjects learn suddenly and their predictions fluctuate perpetually. We suggest that abrupt and unstable learning can be modeled by assuming subjects are conducting in- ference using sequential Monte Carlo sampling with a small number of samples — one, in our simulations. Ensemble behavior resembles exact Bayesian models since, as in particle filters, it averages over many samples. Further, the model is capable of exhibiting sophisticated behaviors like retrospective revaluation at the ensemble level, even given minimally sophisticated individuals that do not track uncertainty in their beliefs over trials.

****

Non-parametric Modeling of Partially Ranked Data

Guy Lebanon, Yi Mao

Statistical models on full and partial rankings of n items are often of limited prac- tical use for large n due to computational consideration. We explore the use of non-parametric models for partially ranked data and derive efficient procedures for their use for large n. The derivations are largely possible throu gh combinatorial and algebraic manipulations based on the lattice of partial rankings. In particular, we demonstrate for the (cid:2)rst time a non-parametric coherent and consistent model capable of ef(cid:2)ciently aggregating partially ranked data of different types.

****

Multiple-Instance Pruning For Learning Efficient Cascade Detectors

Cha Zhang, Paul Viola

Cascade detectors have been shown to operate extremely rapidly, with high accuracy, and have important applications such as face detection. Driven by this success, cascade earning has been an area of active research in recent years. Nevertheless, there are still challenging technical problems during the training process of cascade detectors. In particular, determining the optimal target detection rate for each stage of the cascade remains an unsolved issue. In this paper, we propose the multiple instance pruning (MIP) algorithm for soft cascades. This algorithm computes a set of thresholds which aggressively terminate computation with no reduction in detection rate or increase in false positive rate on the training dataset. The algorithm is based on two key insights: i) examples that are destined to be rejected by the complete classifier can be safely pruned early; ii) face detection is a multiple instance learning problem. The MIP process is fully automatic and requires no assumptions of probability distributions, statistical independence, or ad hoc intermediate rejection targets. Experimental results on the MIT+CMU dataset demonstrate significant performance advantages.

****