Monte Carlo Sampling for Regret Minimization in Extensive Games
Marc Lanctot, Kevin Waugh, Martin Zinkevich, Michael Bowling
Sequential decision-making with multiple agents and imperfect information is commonly modeled as an extensive game. One efficient method for computing Nash equilibria in large, zero-sum, imperfect information games is counterfactual regret minimization (CFR). In the domain of poker, CFR has proven effective, particularly when using a domain-specific augmentation involving chance outcome sampling.
In this paper, we describe a general family of domain independent CFR sample-based algorithms called Monte Carlo counterfactual regret minimization (MCCFR) of which the original and poker-specific versions are special cases. We start by showing that MCCFR performs the same regret updates as CFR on expectation. Then, we introduce two sampling schemes: {\it outcome sampling} and {\it external sampling}, showing that both have bounded overall regret with high probability. Thus, they can compute an approximate equilibrium using self-play. Finally, we prove a new tighter bound on the regret for the original CFR algorithm and relate this new bound to MCCFRs bounds. We show empirically that, although the sample-based algorithms require more iterations, their lower cost per iteration can lead to dramatically faster convergence in various games.
****
DUOL: A Double Updating Approach for Online Learning
Peilin Zhao, Steven Hoi, Rong Jin
In most online learning algorithms, the weights assigned to the misclassified examples (or support vectors) remain unchanged during the entire learning process. This is clearly insufficient since when a new misclassified example is added to the pool of support vectors, we generally expect it to affect the weights for the existing support vectors. In this paper, we propose a new online learning method, termed Double Updating Online Learning", or "DUOL" for short. Instead of only assigning a fixed weight to the misclassified example received in current trial, the proposed online learning algorithm also tries to update the weight for one of the existing support vectors. We show that the mistake bound can be significantly improved by the proposed online learning method. Encouraging experimental results show that the proposed technique is in general considerably more effective than the state-of-the-art online learning algorithms."
****
Compressed Least-Squares Regression
Odalric Maillard, Rémi Munos
We consider the problem of learning, from K input data, a regression function in a function space of high dimension N using projections onto a random subspace of lower dimension M. From any linear approximation algorithm using empirical risk minimization (possibly penalized), we provide bounds on the excess risk of the estimate computed in the projected subspace (compressed domain) in terms of the excess risk of the estimate built in the high-dimensional space (initial domain). We apply the analysis to the ordinary Least-Squares regression and show that by choosing M=O(\sqrt{K}), the estimation error (for the quadratic loss) of the ``Compressed Least Squares Regression is O(1/\sqrt{K}) up to logarithmic factors. We also discuss the numerical complexity of several algorithms (both in initial and compressed domains) as a function of N, K, and M.
****
White Functionals for Anomaly Detection in Dynamical Systems
Marco Cuturi, Jean-philippe Vert, Alexandre D'aspremont
We propose new methodologies to detect anomalies in discrete-time processes taking values in a set. The method is based on the inference of functionals whose evaluations on successive states visited by the process have low autocorrelations. Deviations from this behavior are used to flag anomalies. The candidate functionals are estimated in a subset of a reproducing kernel Hilbert space associated with the set where the process takes values. We provide experimental results which show that these techniques compare favorably with other algorithms.
****
Learning models of object structure
Joseph Schlecht, Kobus Barnard

We present an approach for learning stochastic geometric models of object categories from single view images. We focus here on models expressible as a spatially contiguous assemblage of blocks. Model topologies are learned across groups of images, and one or more such topologies is linked to an object category (e.g. chairs). Fitting learned topologies to an image can be used to identify the object class, as well as detail its geometry. The latter goes beyond labeling objects, as it provides the geometric structure of particular instances. We learn the models using joint statistical inference over structure parameters, camera parameters, and instance parameters. These produce an image likelihood through a statistical imaging model. We use trans-dimensional sampling to explore topology hypotheses, and alternate between Metropolis-Hastings and stochastic dynamics to explore instance parameters. Experiments on images of furniture objects such as tables and chairs suggest that this is an effective approach for learning models that encode simple representations of category geometry and the statistics thereof, and support inferring both category and geometry on held out single view images.

****

Neurometric function analysis of population codes

Philipp Berens, Sebastian Gerwinn, Alexander Ecker, Matthias Bethge

The relative merits of different population coding schemes have mostly been analyzed in the framework of stimulus reconstruction using Fisher Information. Here, we consider the case of stimulus discrimination in a two alternative forced choice paradigm and compute neurometric functions in terms of the minimal discrimination error and the Jensen-Shannon information to study neural population codes. We first explore the relationship between minimum discrimination error, Jensen-Shannon Information and Fisher Information and show that the discrimination framework is more informative about the coding accuracy than Fisher Information as it defines an error for any pair of possible stimuli. In particular, it includes Fisher Information as a special case. Second, we use the framework to study population codes of angular variables. Specifically, we assess the impact of different noise correlations structures on coding accuracy in long versus short decoding time windows. That is, for long time window we use the common Gaussian noise approximation. To address the case of short time windows we analyze the Ising model with identical noise correlation structure. In this way, we provide a new rigorous framework for assessing the functional consequences of noise correlation structures for the representational accuracy of neural population codes that is in particular applicable to short-time population coding.

****

Slow, Decorrelated Features for Pretraining Complex Cell-like Networks

Yoshua Bengio, James Bergstra

We introduce a new type of neural network activation function based on recent physiological rate models for complex cells in visual area V1. A single-hidden-layer neural network of this kind of model achieves 1.5% error on MNIST. We also introduce an existing criterion for learning slow, decorrelated features as a pretraining strategy for image models. This pretraining strategy results in orientation-selective features, similar to the receptive fields of complex cells. With this pretraining, the same single-hidden-layer model achieves better generalization error, even though the pretraining sample distribution is very different from the fine-tuning distribution. To implement this pretraining strategy, we derive a fast algorithm for online learning of decorrelated features such that each iteration of the algorithm runs in linear time with respect to the number of features.

****

Code-specific policy gradient rules for spiking neurons

Henning Sprekeler, Guillaume Hennequin, Wulfram Gerstner

Although it is widely believed that reinforcement learning is a suitable tool for describing behavioral learning, the mechanisms by which it can be implemented in networks of spiking neurons are not fully understood. Here, we show that different learning rules emerge from a policy gradient approach depending on which features of the spike trains are assumed to influence the reward signals, i.e., d

epending on which neural code is in effect. We use the framework of Williams (19
92) to derive learning rules for arbitrary neural codes. For illustration, we pr
esent policy-gradient rules for three different example codes - a spike count c
ode, a spike timing code and the most general ``full spike train code - and test
them on simple model problems. In addition to classical synaptic learning, we d
erive learning rules for intrinsic parameters that control the excitability of t
he neuron. The spike count learning rule has structural similarities with establ
ished Bienenstock-Cooper-Munro rules. If the distribution of the relevant spike
train features belongs to the natural exponential family, the learning rules ha
ve a characteristic shape that raises interesting prediction problems.
****

Positive Semidefinite Metric Learning with Boosting
Chunhua Shen, Junae Kim, Lei Wang, Anton Hengel
The learning of appropriate distance metrics is a critical problem in classifica
tion. In this work, we propose a boosting-based technique, termed BoostMetric, f
or learning a Mahalanobis distance metric. One of the primary difficulties in le
arning such a metric is to ensure that the Mahalanobis matrix remains positive s
emidefinite. Semidefinite programming is sometimes used to enforce this constrai
nt, but does not scale well. BoostMetric is instead based on a key observation t
hat any positive semidefinite matrix can be decomposed into a linear positive co
mbination of trace-one rank-one matrices. BoostMetric thus uses rank-one positi
ve semidefinite matrices as weak learners within an efficient and scalable boost
ing-based learning process. The resulting method is easy to implement, does not
require tuning, and can accommodate various types of constraints. Experiments o
n various datasets show that the proposed algorithm compares favorably to those
state-of-the-art methods in terms of classification accuracy and running time.
****

Spatial Normalized Gamma Processes
Vinayak Rao, Yee Teh
Dependent Dirichlet processes (DPs) are dependent sets of random measures, each
being marginally Dirichlet process distributed. They are used in Bayesian nonpar
ametric models when the usual exchangebility assumption does not hold. We propos
e a simple and general framework to construct dependent DPs by marginalizing and
normalizing a single gamma process over an extended space. The result is a set
of DPs, each located at a point in a space such that neighboring DPs are more d
ependent. We describe Markov chain Monte Carlo inference, involving the typical
Gibbs sampling and three different Metropolis-Hastings proposals to speed up con
vergence. We report an empirical study of convergence speeds on a synthetic data
set and demonstrate an application of the model to topic modeling through time.
****

Entropic Graph Regularization in Non-Parametric Semi-Supervised Classification
Amarnag Subramanya, Jeff A. Bilmes
We prove certain theoretical properties of a graph-regularized transductive lear
ning objective that is based on minimizing a Kullback-Leibler divergence based l
oss. These include showing that the iterative alternating minimization procedure
used to minimize the objective converges to the correct solution and deriving a
test for convergence. We also propose a graph node ordering algorithm that is c
ache cognizant and leads to a linear speedup in parallel computations. This ensu
res that the algorithm scales to large data sets. By making use of empirical eva
luation on the TIMIT and Switchboard I corpora, we show this approach is able to
out-perform other state-of-the-art SSL approaches. In one instance, we solve a
problem on a 120 million node graph.
****

Replacing supervised classification learning by Slow Feature Analysis in spiking
neural networks
Stefan Klampfl, Wolfgang Maass
Many models for computations in recurrent networks of neurons assume that the ne
twork state moves from some initial state to some fixed point attractor or limit
cycle that represents the output of the computation. However experimental data
show that in response to a sensory stimulus the network state moves from its ini

tial state through a trajectory of network states and eventually returns to the initial state, without reaching an attractor or limit cycle in between. This type of network response, where salient information about external stimuli is encoded in characteristic trajectories of continuously varying network states, raises the question how a neural system could compute with such code, and arrive for example at a temporally stable classification of the external stimulus. We show that a known unsupervised learning algorithm, Slow Feature Analysis (SFA), could be an important ingredient for extracting stable information from these network trajectories. In fact, if sensory stimuli are more often followed by another stimulus from the same class than by a stimulus from another class, SFA approaches the classification capability of Fishers Linear Discriminant (FLD), a powerful algorithm for supervised learning. We apply this principle to simulated cortical microcircuits, and show that it enables readout neurons to learn discrimination of spoken digits and detection of repeating firing patterns within a stream of spike trains with the same firing statistics, without requiring any supervision for learning.

****

Free energy score space

Alessandro Perina, Marco Cristani, Umberto Castellani, Vittorio Murino, Nebojsa Jojic

Score functions induced by generative models extract fixed-dimension feature vectors from different-length data observations by subsuming the process of data generation, projecting them in highly informative spaces called score spaces. In this way, standard discriminative classifiers are proved to achieve higher performances than a solely generative or discriminative approach. In this paper, we present a novel score space that exploits the free energy associated to a generative model through a score function. This function aims at capturing both the uncertainty of the model learning and ``local compliance of data observations with respect to the generative process. Theoretical justifications and convincing comparative classification results on various generative models prove the goodness of the proposed strategy.

****

Multi-Label Prediction via Sparse Infinite CCA

Piyush Rai, Hal Daume

Canonical Correlation Analysis (CCA) is a useful technique for modeling dependencies between two (or more) sets of variables. Building upon the recently suggested probabilistic interpretation of CCA, we propose a nonparametric, fully Bayesian framework that can automatically select the number of correlation components, and effectively capture the sparsity underlying the projections. In addition, given (partially) labeled data, our algorithm can also be used as a (semi)supervised dimensionality reduction technique, and can be applied to learn useful predictive features in the context of learning a set of related tasks. Experimental results demonstrate the efficacy of the proposed approach for both CCA as a stand-alone problem, and when applied to multi-label prediction.

****

Fast subtree kernels on graphs

Nino Shervashidze, Karsten Borgwardt

In this article, we propose fast subtree kernels on graphs. On graphs with n nodes and m edges and maximum degree d, these kernels comparing subtrees of height h can be computed in O(mh), whereas the classic subtree kernel by Ramon & G¨artner scales as O(n24dh). Key to this ef■ciency is the observation that the Weisfeiler-Lehman test of isomorphism from graph theory elegantly computes a subtree kernel as a byproduct. Our fast subtree kernels can deal with labeled graphs, scale up easily to large graphs and outperform state-of-the-art graph ker- nels on several classi■cation benchmark datasets in terms of accuracy and runtime.

****

Directed Regression

Yi-hao Kao, Benjamin Roy, Xiang Yan

When used to guide decisions, linear regression analysis typically involves estimation of regression coefficients via ordinary least squares and their subsequen

t use to make decisions. When there are multiple response variables and features do not perfectly capture their relationships, it is beneficial to account for the decision objective when computing regression coefficients. Empirical optimization does so but sacrifices performance when features are well-chosen or training data are insufficient. We propose directed regression, an efficient algorithm that combines merits of ordinary least squares and empirical optimization. We demonstrate through a computational study that directed regression can generate significant performance gains over either alternative. We also develop a theory that motivates the algorithm.
****

Nonparametric Greedy Algorithms for the Sparse Learning Problem
Han Liu, Xi Chen
This paper studies the forward greedy strategy in sparse nonparametric regression. For additive models, we propose an algorithm called additive forward regression; for general multivariate regression, we propose an algorithm called generalized forward regression. Both of them simultaneously conduct estimation and variable selection in nonparametric settings for the high dimensional sparse learning problem. Our main emphasis is empirical: on both simulated and real data, these two simple greedy methods can clearly outperform several state-of-the-art competitors, including the LASSO, a nonparametric version of the LASSO called the sparse additive model (SpAM) and a recently proposed adaptive parametric forward-backward algorithm called the Foba. Some theoretical justifications are also provided.
****

Rethinking LDA: Why Priors Matter
Hanna Wallach, David Mimno, Andrew McCallum
Implementations of topic models typically use symmetric Dirichlet priors with fixed concentration parameters, with the implicit assumption that such smoothing parameters" have little practical effect. In this paper, we explore several classes of structured priors for topic models. We find that an asymmetric Dirichlet prior over the document-topic distributions has substantial advantages over a symmetric prior, while an asymmetric prior over the topic-word distributions provides no real benefit. Approximation of this prior structure through simple, efficient hyperparameter optimization steps is sufficient to achieve these performance gains. The prior structure we advocate substantially increases the robustness of topic models to variations in the number of topics and to the highly skewed word frequency distributions common in natural language. Since this prior structure can be implemented using efficient algorithms that add negligible cost beyond standard inference techniques, we recommend it as a new standard for topic modeling."
****

Augmenting Feature-driven fMRI Analyses: Semi-supervised learning and resting state activity
Andreas Bartels, Matthew Blaschko, Jacquelyn Shelton
Resting state activity is brain activation that arises in the absence of any task, and is usually measured in awake subjects during prolonged fMRI scanning sessions where the only instruction given is to close the eyes and do nothing. It has been recognized in recent years that resting state activity is implicated in a wide variety of brain function. While certain networks of brain areas have different levels of activation at rest and during a task, there is nevertheless significant similarity between activations in the two cases. This suggests that recordings of resting state activity can be used as a source of unlabeled data to augment discriminative regression techniques in a semi-supervised setting. We evaluate this setting empirically yielding three main results: (i) regression tends to be improved by the use of Laplacian regularization even when no additional unlabeled data are available, (ii) resting state data may have a similar marginal distribution to that recorded during the execution of a visual processing task reinforcing the hypothesis that these conditions have similar types of activation, and (iii) this source of information can be broadly exploited to improve the robustness of empirical inference in fMRI studies, an inherently data poor do

main.
****
Differential Use of Implicit Negative Evidence in Generative and Discriminative Language Learning
Anne Hsu, Thomas Griffiths
A classic debate in cognitive science revolves around understanding how children learn complex linguistic rules, such as those governing restrictions on verb alternations, without negative evidence. Traditionally, formal learnability arguments have been used to claim that such learning is impossible without the aid of innate language-specific knowledge. However, recently, researchers have shown that statistical models are capable of learning complex rules from only positive evidence. These two kinds of learnability analyses differ in their assumptions about the role of the distribution from which linguistic input is generated. The former analyses assume that learners seek to identify grammatical sentences in a way that is robust to the distribution from which the sentences are generated, analogous to discriminative approaches in machine learning. The latter assume that learners are trying to estimate a generative model, with sentences being sampled from that model. We show that these two learning approaches differ in their use of implicit negative evidence -- the absence of a sentence -- when learning verb alternations, and demonstrate that human learners can produce results consistent with the predictions of both approaches, depending on the context in which the learning problem is presented.
****
Non-stationary continuous dynamic Bayesian networks
Marco Grzegorczyk, Dirk Husmeier
Dynamic Bayesian networks have been applied widely to reconstruct the structure of regulatory processes from time series data. The standard approach is based on the assumption of a homogeneous Markov chain, which is not valid in many real-world scenarios. Recent research efforts addressing this shortcoming have considered undirected graphs, directed graphs for discretized data, or over-flexible models that lack any information sharing between time series segments. In the present article, we propose a non-stationary dynamic Bayesian network for continuous data, in which parameters are allowed to vary between segments, and in which a common network structure provides essential information sharing across segments. Our model is based on a Bayesian change-point process, and we apply a variant of the allocation sampler of Nobile and Fearnside to infer the number and location of the change-points.
****
Learning Brain Connectivity of Alzheimer's Disease from Neuroimaging Data
Shuai Huang, Jing Li, Liang Sun, Jun Liu, Teresa Wu, Kewei Chen, Adam Fleisher, Eric Reiman, Jieping Ye
Recent advances in neuroimaging techniques provide great potentials for effective diagnosis of Alzheimer’s disease (AD), the most common form of dementia. Previous studies have shown that AD is closely related to alternation in the functional brain network, i.e., the functional connectivity among different brain regions. In this paper, we consider the problem of learning functional brain connectivity from neuroimaging, which holds great promise for identifying image-based markers used to distinguish Normal Controls (NC), patients with Mild Cognitive Impairment (MCI), and patients with AD. More specifically, we study sparse inverse covariance estimation (SICE), also known as exploratory Gaussian graphical models, for brain connectivity modeling. In particular, we apply SICE to learn and analyze functional brain connectivity patterns from different subject groups, based on a key property of SICE, called the “monotone property” we established in this paper. Our experimental results on neuroimaging PET data of 42 AD, 116 MCI, and 67 NC subjects reveal several interesting connectivity patterns consistent with literature findings, and also some new patterns that can help the knowledge discovery of AD.
****
Potential-Based Agnostic Boosting
Varun Kanade, Adam Kalai

We prove strong noise-tolerance properties of a potential-based boosting algorithm, similar to MadaBoost (Domingo and Watanabe, 2000) and SmoothBoost (Servedio, 2003). Our analysis is in the agnostic framework of Kearns, Schapire and Sellie (1994), giving polynomial-time guarantees in presence of arbitrary noise. A remarkable feature of our algorithm is that it can be implemented without reweighting examples, by randomly relabeling them instead. Our boosting theorem gives, as easy corollaries, alternative derivations of two recent non-trivial results in computational learning theory: agnostically learning decision trees (Gopalan et al, 2008) and agnostically learning halfspaces (Kalai et al, 2005). Experiments suggest that the algorithm performs similarly to Madaboost.
****

Gaussian process regression with Student-t likelihood
Jarno Vanhatalo, Pasi Jylänki, Aki Vehtari
In the Gaussian process regression the observation model is commonly assumed to be Gaussian, which is convenient in computational perspective. However, the drawback is that the predictive accuracy of the model can be significantly compromised if the observations are contaminated by outliers. A robust observation model, such as the Student-t distribution, reduces the influence of outlying observations and improves the predictions. The problem, however, is the analytically intractable inference. In this work, we discuss the properties of a Gaussian process regression model with the Student-t likelihood and utilize the Laplace approximation for approximate inference. We compare our approach to a variational approximation and a Markov chain Monte Carlo scheme, which utilize the commonly used scale mixture representation of the Student-t distribution.
****

Bilinear classifiers for visual recognition
Hamed Pirsiavash, Deva Ramanan, Charless Fowlkes
We describe an algorithm for learning bilinear SVMs. Bilinear classifiers are a discriminative variant of bilinear models, which capture the dependence of data on multiple factors. Such models are particularly appropriate for visual data that is better represented as a matrix or tensor, rather than a vector. Matrix encodings allow for more natural regularization through rank restriction. For example, a rank-one scanning-window classifier yields a separable filter. Low-rank models have fewer parameters and so are easier to regularize and faster to score at run-time. We learn low-rank models with bilinear classifiers. We also use bilinear classifiers for transfer learning by sharing linear factors between different classification tasks. Bilinear classifiers are trained with biconvex programs. Such programs are optimized with coordinate descent, where each coordinate step requires solving a convex program - in our case, we use a standard off-the-shelf SVM solver. We demonstrate bilinear SVMs on difficult problems of people detection in video sequences and action classification of video sequences, achieving state-of-the-art results in both.
****

Zero-shot Learning with Semantic Output Codes
Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, Tom M. Mitchell
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.
****

Learning Label Embeddings for Nearest-Neighbor Multi-class Classification with an Application to Speech Recognition
Natasha Singh-miller, Michael Collins
We consider the problem of using nearest neighbor methods to provide a conditional probability estimate, P(y|a), when the number of labels y is large and the labels share some underlying structure. We propose a method for learning error-correcting output codes (ECOCs) to model the similarity between labels within a nearest neighbor framework. The learned ECOCs and nearest neighbor information are used to provide conditional probability estimates. We apply these estimates to the problem of acoustic modeling for speech recognition. We demonstrate an absolu

te reduction in word error rate (WER) of 0.9% (a 2.5% relative reduction in WER) on a lecture recognition task over a state-of-the-art baseline GMM model.
****

Semi-Supervised Learning in Gigantic Image Collections
Rob Fergus, Yair Weiss, Antonio Torralba
With the advent of the Internet it is now possible to collect hundreds of millions of images. These images come with varying degrees of label information. Clean labels can be manually obtained on a small fraction,noisy labels may be extracted automatically from surrounding text, while for most images there are no labels at all. Semi-supervised learning is a principled framework for combining these different label sources. However, it scales polynomially with the number of images, making it impractical for use on gigantic collections with hundreds of millions of images and thousands of classes. In this paper we show how to utilize recent results in machine learning to obtain highly efficient approximations for semi-supervised learning that are linear in the number of images. Specifically, we use the convergence of the eigenvectors of the normalized graph Laplacian to eigenfunctions of weighted Laplace-Beltrami operators. We combine this with a label sharing framework obtained from Wordnet to propagate label information to classes lacking manual annotations. Our algorithm enables us to apply semi-supervised learning to a database of 80 million images with 74 thousand classes.
****

Sensitivity analysis in HMMs with application to likelihood maximization
Pierre-arnaud Coquelin, Romain Deguest, Rémi Munos
This paper considers a sensitivity analysis in Hidden Markov Models with continuous state and observation spaces. We propose an Infinitesimal Perturbation Analysis (IPA) on the filtering distribution with respect to some parameters of the model. We describe a methodology for using any algorithm that estimates the filtering density, such as Sequential Monte Carlo methods, to design an algorithm that estimates its gradient. The resulting IPA estimator is proven to be asymptotically unbiased, consistent and has computational complexity linear in the number of particles. We consider an application of this analysis to the problem of identifying unknown parameters of the model given a sequence of observations. We derive an IPA estimator for the gradient of the log-likelihood, which may be used in a gradient method for the purpose of likelihood maximization. We illustrate the method with several numerical experiments.
****

Who’s Doing What: Joint Modeling of Names and Verbs for Simultaneous Face and Pose Annotation
Jie Luo, Barbara Caputo, Vittorio Ferrari
Given a corpus of news items consisting of images accompanied by text captions, we want to find out `whos doing what, i.e. associate names and action verbs in the captions to the face and body pose of the persons in the images. We present a joint model for simultaneously solving the image-caption correspondences and learning visual appearance models for the face and pose classes occurring in the corpus. These models can then be used to recognize people and actions in novel images without captions. We demonstrate experimentally that our jointface and pose model solves the correspondence problem better than earlier models covering only the face, and that it can perform recognition of new uncaptioned images.
****

Streaming Pointwise Mutual Information
Benjamin Durme, Ashwin Lall
Recent work has led to the ability to perform space ef■cient, approximate counting over large vocabularies in a streaming context. Motivated by the existence of data structures of this type, we explore the computation of associativity scores, other- wise known as pointwise mutual information (PMI), in a streaming context. We give theoretical bounds showing the impracticality of perfect online PMI compu- tation, and detail an algorithm with high expected accuracy. Experiments on news articles show our approach gives high accuracy on real world data.
****

Nonparametric Bayesian Models for Unsupervised Event Coreference Resolution
Cosmin Bejan, Matthew Titsworth, Andrew Hickl, Sanda Harabagiu
We present a sequence of unsupervised, nonparametric Bayesian models for cluster ing complex linguistic objects. In this approach, we consider a potentially inf inite number of features and categorical outcomes. We evaluate these models for the task of within- and cross-document event coreference on two corpora. All t he models we investigated show significant improvements when compared against an existing baseline for this task.
****
A Stochastic approximation method for inference in probabilistic graphical model s
Peter Carbonetto, Matthew King, Firas Hamze
We describe a new algorithmic framework for inference in probabilistic models, a nd apply it to inference for latent Dirichlet allocation. Our framework adopts t he methodology of variational inference, but unlike existing variational methods such as mean field and expectation propagation it is not restricted to tractabl e classes of approximating distributions. Our approach can also be viewed as a s equential Monte Carlo (SMC) method, but unlike existing SMC methods there is no need to design the artificial sequence of distributions. Notably, our framework offers a principled means to exchange the variance of an importance sampling est imate for the bias incurred through variational approximation. Experiments on a challenging inference problem in population genetics demonstrate improvements in stability and accuracy over existing methods, and at a comparable cost.
****
Factor Modeling for Advertisement Targeting
Ye Chen, Michael Kapralov, John Canny, Dmitry Pavlov
We adapt a probabilistic latent variable model, namely GaP (Gamma-Poisson), to a d targeting in the contexts of sponsored search (SS) and behaviorally targeted (BT) display advertising. We also approach the important problem of ad positional bias by formulating a one-latent-dimension GaP factorization. Learning from cli ck-through data is intrinsically large scale, even more so for ads. We scale up the algorithm to terabytes of real-world SS and BT data that contains hundreds o f millions of users and hundreds of thousands of features, by leveraging the sca lability characteristics of the algorithm and the inherent structure of the prob lem including data sparsity and locality. Specifically, we demonstrate two somew hat orthogonal philosophies of scaling algorithms to large-scale problems, throu gh the SS and BT implementations, respectively. Finally, we report the experimen tal results using Yahoos vast datasets, and show that our approach substantially outperform the state-of-the-art methods in prediction accuracy. For BT in parti cular, the ROC area achieved by GaP is exceeding 0.95, while one prior approach using Poisson regression yielded 0.83. For computational performance, we compare a single-node sparse implementation with a parallel implementation using Hadoop MapReduce, the results are counterintuitive yet quite interesting. We therefore provide insights into the underlying principles of large-scale learning.
****
Sparse Estimation Using General Likelihoods and Non-Factorial Priors
David Wipf, Srikantan Nagarajan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Modeling the spacing effect in sequential category learning
Hongjing Lu, Matthew Weiden, Alan L. Yuille
We develop a Bayesian sequential model for category learning. The sequential mod el updates two category parameters, the mean and the variance, over time. We def ine conjugate temporal priors to enable closed form solutions to be obtained. Th is model can be easily extended to supervised and unsupervised learning involvin g multiple categories. To model the spacing effect, we introduce a generic prior in the temporal updating stage to capture a learning preference, namely, less c

hange for repetition and more change for variation. Finally, we show how this ap
proach can be generalized to efficiently performmodel selection to decide whethe
r observations are from one or multiple categories.
****

An Infinite Factor Model Hierarchy Via a Noisy-Or Mechanism
Douglas Eck, Yoshua Bengio, Aaron C. Courville
The Indian Buffet Process is a Bayesian nonparametric approach that models objec
ts as arising from an infinite number of latent factors. Here we extend the late
nt factor model framework to two or more unbounded layers of latent factors. Fro
m a generative perspective, each layer defines a conditional \emph{factorial} pr
ior distribution over the binary latent variables of the layer below via a noisy
-or mechanism. We explore the properties of the model with two empirical studies
, one digit recognition task and one music tag data experiment.
****

Large Scale Nonparametric Bayesian Inference: Data Parallelisation in the Indian
 Buffet Process
Finale Doshi-velez, Shakir Mohamed, Zoubin Ghahramani, David Knowles
Nonparametric Bayesian models provide a framework for flexible probabilistic mod
elling of complex datasets. Unfortunately, Bayesian inference methods often requ
ire high-dimensional averages and can be slow to compute, especially with the po
tentially unbounded representations associated with nonparametric models. We add
ress the challenge of scaling nonparametric Bayesian inference to the increasing
ly large datasets found in real-world applications, focusing on the case of para
llelising inference in the Indian Buffet Process (IBP). Our approach divides a
large data set between multiple processors. The processors use message passing
to compute likelihoods in an asynchronous, distributed fashion and to propagate
statistics about the global Bayesian posterior. This novel MCMC sampler is the
first parallel inference scheme for IBP-based models, scaling to datasets orders
 of magnitude larger than had previously been possible.
****

Matrix Completion from Power-Law Distributed Samples
Raghu Meka, Prateek Jain, Inderjit Dhillon
The low-rank matrix completion problem is a fundamental problem with many import
ant applications. Recently, Candes & Recht, Keshavan et al. and Candes & Tao obt
ained the first non-trivial theoretical results for the problem assuming that th
e observed entries are sampled uniformly at random. Unfortunately, most real-wor
ld datasets do not satisfy this assumption, but instead exhibit power-law distri
buted samples. In this paper, we propose a graph theoretic approach to matrix co
mpletion that solves the problem for more realistic sampling models. Our method
is easier to analyze than previous methods with the analysis reducing to computi
ng the threshold for complete cascades in random graphs, a problem of independen
t interest. By analyzing the graph theoretic problem, we show that our method ac
hieves exact recovery when the observed entries are sampled from the Chung-Lu-Vu
 model, which can generate power-law distributed graphs. We also hypothesize tha
t our algorithm solves the matrix completion problem from an optimal number of e
ntries for the popular preferential attachment model and provide strong empirica
l evidence for the claim. Furthermore, our method is easier to implement and is
substantially faster than existing methods. We demonstrate the effectiveness of
our method on examples when the low-rank matrix is sampled according to the prev
alent random graph models for complex networks and also on the Netflix challenge
 dataset.
****

Filtering Abstract Senses From Image Search Results
Kate Saenko, Trevor Darrell
We propose an unsupervised method that, given a word, automatically selects non-
abstract senses of that word from an online ontology and generates images depict
ing the corresponding entities. When faced with the task of learning a visual mo
del based only on the name of an object, a common approach is to find images on
the web that are associated with the object name, and then train a visual classi
fier from the search result. As words are generally polysemous, this approach ca

n lead to relatively noisy models if many examples due to outlier senses are add ed to the model. We argue that images associated with an abstract word sense sho uld be excluded when training a visual classifier to learn a model of a physical object. While image clustering can group together visually coherent sets of ret urned images, it can be difficult to distinguish whether an image cluster relate s to a desired object or to an abstract sense of the word. We propose a method t hat uses both image features and the text associated with the images to relate l atent topics to particular senses. Our model does not require any human supervis ion, and takes as input only the name of an object category. We show results of retrieving concrete-sense images in two available multimodal, multi-sense databa ses, as well as experiment with object classifiers trained on concrete-sense ima ges returned by our method for a set of ten common office objects.

****

Heavy-Tailed Symmetric Stochastic Neighbor Embedding

Zhirong Yang, Irwin King, Zenglin Xu, Erkki Oja

Stochastic Neighbor Embedding (SNE) has shown to be quite promising for data vis ualization. Currently, the most popular implementation, t-SNE, is restricted to a particular Student t-distribution as its embedding distribution. Moreover, it uses a gradient descent algorithm that may require users to tune parameters suc h as the learning step size, momentum, etc., in finding its optimum. In this pap er, we propose the Heavy-tailed Symmetric Stochastic Neighbor Embedding (HSSNE) method, which is a generalization of the t-SNE to accommodate various heavy-tail ed embedding similarity functions. With this generalization, we are presented wi th two difficulties. The first is how to select the best embedding similarity a mong all heavy-tailed functions and the second is how to optimize the objective function once the heave-tailed function has been selected. Our contributions the n are: (1) we point out that various heavy-tailed embedding similarities can be characterized by their negative score functions. Based on this finding, we prese nt a parameterized subset of similarity functions for choosing the best tail-hea viness for HSSNE; (2) we present a fixed-point optimization algorithm that can b e applied to all heavy-tailed functions and does not require the user to set any parameters; and (3) we present two empirical studies, one for unsupervised visu alization showing that our optimization algorithm runs as fast and as good as th e best known t-SNE implementation and the other for semi-supervised visualizatio n showing quantitative superiority using the homogeneity measure as well as qual itative advantage in cluster separation over t-SNE.

****

Which graphical models are difficult to learn?

Andrea Montanari, Jose Pereira

We consider the problem of learning the structure of Ising models (pairwise bina ry Markov random fields) from i.i.d. samples. While several methods have been pr oposed to accomplish this task, their relative merits and limitations remain som ewhat obscure. By analyzing a number of concrete examples, we show that low-comp lexity algorithms systematically fail when the Markov random field develops lo ng-range correlations. More precisely, this phenomenon appears to be related to the Ising model phase transition (although it does not coincide with it).

****

Information-theoretic lower bounds on the oracle complexity of convex optimizati on

Alekh Agarwal, Martin J. Wainwright, Peter Bartlett, Pradeep Ravikumar

Despite the large amount of literature on upper bounds on complexity of convex a nalysis, surprisingly little is known about the fundamental hardness of these pr oblems. The extensive use of convex optimization in machine learning and statist ics makes such an understanding critical to understand fundamental computational limits of learning and estimation. In this paper, we study the complexity of st ochastic convex optimization in an oracle model of computation. We improve upon known results and obtain tight minimax complexity estimates for some function cl asses. We also discuss implications of these results to the understanding the in herent complexity of large-scale learning and estimation problems.

****

Linear-time Algorithms for Pairwise Statistical Problems
Parikshit Ram, Dongryeol Lee, William March, Alexander Gray
Several key computational bottlenecks in machine learning involve pairwise distance computations, including all-nearest-neighbors (finding the nearest neighbor(s) for each point, e.g. in manifold learning) and kernel summations (e.g. in kernel density estimation or kernel machines). We consider the general, bichromatic case for these problems, in addition to the scientific problem of N-body potential calculation. In this paper we show for the first time O(N) worst case run times for practical algorithms for these problems based on the cover tree data structure (Beygelzimer, Kakade, Langford, 2006).
****

From PAC-Bayes Bounds to KL Regularization
Pascal Germain, Alexandre Lacasse, Mario Marchand, Sara Shanian, François Laviolette
We show that convex KL-regularized objective functions are obtained from a PAC-Bayes risk bound when using convex loss functions for the stochastic Gibbs classifier that upper-bound the standard zero-one loss used for the weighted majority vote. By restricting ourselves to a class of posteriors, that we call quasi uniform, we propose a simple coordinate descent learning algorithm to minimize the proposed KL-regularized cost function. We show that standard ellp-regularized objective functions currently used, such as ridge regression and ellp-regularized boosting, are obtained from a relaxation of the KL divergence between the quasi uniform posterior and the uniform prior. We present numerical experiments where the proposed learning algorithm generally outperforms ridge regression and AdaBoost.
****

On Stochastic and Worst-case Models for Investing
Elad Hazan, Satyen Kale
In practice, most investing is done assuming a probabilistic model of stock price returns known as the Geometric Brownian Motion (GBM). While it is often an acceptable approximation, the GBM model is not always valid empirically. This motivates a worst-case approach to investing, called universal portfolio management, where the objective is to maximize wealth relative to the wealth earned by the best fixed portfolio in hindsight. In this paper we tie the two approaches, and design an investment strategy which is universal in the worst-case, and yet capable of exploiting the mostly valid GBM model. Our method is based on new and improved regret bounds for online convex optimization with exp-concave loss functions.
****

Structured output regression for detection with partial truncation
Andrea Vedaldi, Andrew Zisserman
We develop a structured output model for object category detection that explicitly accounts for alignment, multiple aspects and partial truncation in both training and inference. The model is formulated as large margin learning with latent variables and slack rescaling, and both training and inference are computationally efficient. We make the following contributions: (i) we note that extending the Structured Output Regression formulation of Blaschko and Lampert (ECCV 2008) to include a bias term significantly improves performance; (ii) that alignment (to account for small rotations and anisotropic scalings) can be included as a latent variable and efficiently determined and implemented; (iii) that the latent variable extends to multiple aspects (e.g. left facing, right facing, front) with the same formulation; and (iv), most significantly for performance, that truncated and truncated instances can be included in both training and inference with an explicit truncation mask. We demonstrate the method by training and testing on the PASCAL VOC 2007 data set -- training includes the truncated examples, and in testing object instances are detected at multiple scales, alignments, and with significant truncations.
****

On Invariance in Hierarchical Models
Jake Bouvrie, Lorenzo Rosasco, Tomaso Poggio

A goal of central importance in the study of hierarchical models for object recognition -- and indeed the visual cortex -- is that of understanding quantitatively the trade-off between invariance and selectivity, and how invariance and discrimination properties contribute towards providing an improved representation useful for learning from data. In this work we provide a general group-theoretic framework for characterizing and understanding invariance in a family of hierarchical models. We show that by taking an algebraic perspective, one can provide a concise set of conditions which must be met to establish invariance, as well as a constructive prescription for meeting those conditions. Analyses in specific cases of particular relevance to computer vision and text processing are given, yielding insight into how and when invariance can be achieved. We find that the minimal sets of transformations intrinsic to the hierarchical model needed to support a particular invariance can be clearly described, thereby encouraging efficient computational implementations.
****

Submanifold density estimation
Arkadas Ozakin, Alexander Gray
Kernel density estimation is the most widely-used practical method for accurate nonparametric density estimation. However, long-standing worst-case theoretical results showing that its performance worsens exponentially with the dimension of the data have quashed its application to modern high-dimensional datasets for decades. In practice, it has been recognized that often such data have a much lower-dimensional intrinsic structure. We propose a small modification to kernel density estimation for estimating probability density functions on Riemannian submanifolds of Euclidean space. Using ideas from Riemannian geometry, we prove the consistency of this modified estimator and show that the convergence rate is determined by the intrinsic dimension of the submanifold. We conclude with empirical results demonstrating the behavior predicted by our theory.
****

Nonlinear Learning using Local Coordinate Coding
Kai Yu, Tong Zhang, Yihong Gong
This paper introduces a new method for semi-supervised learning on high dimensional nonlinear manifolds, which includes a phase of unsupervised basis learning and a phase of supervised function learning. The learned bases provide a set of anchor points to form a local coordinate system, such that each data point x on the manifold can be locally approximated by a linear combination of its nearby anchor points, and the linear weights become its local coordinate coding. We show that a high dimensional nonlinear function can be approximated by a global linear function with respect to this coding scheme, and the approximation quality is ensured by the locality of such coding. The method turns a difficult nonlinear learning problem into a simple global linear learning problem, which overcomes some drawbacks of traditional local learning methods.
****

No evidence for active sparsification in the visual cortex
Pietro Berkes, Ben White, Jozsef Fiser
The proposal that cortical activity in the visual cortex is optimized for sparse neural activity is one of the most established ideas in computational neuroscience. However, direct experimental evidence for optimal sparse coding remains inconclusive, mostly due to the lack of reference values on which to judge the measured sparseness. Here we analyze neural responses to natural movies in the primary visual cortex of ferrets at different stages of development, and of rats while awake and under different levels of anesthesia. In contrast with prediction from a sparse coding model, our data shows that population and lifetime sparseness decrease with visual experience, and increase from the awake to anesthetized state. These results suggest that the representation in the primary visual cortex is not actively optimized to maximize sparseness.
****

Distribution Matching for Transduction
Novi Quadrianto, James Petterson, Alex Smola
Many transductive inference algorithms assume that distributions over training a

nd test estimates should be related, e.g. by providing a large margin of separat ion on both sets. We use this idea to design a transduction algorithm which can be used without modification for classification, regression, and structured esti mation. At its heart we exploit the fact that for a good learner the distributio ns over the outputs on training and test sets should match. This is a classical two-sample problem which can be solved efficiently in its most general form by u sing distance measures in Hilbert Space. It turns out that a number of existing heuristics can be viewed as special cases of our approach.
****

Canonical Time Warping for Alignment of Human Behavior
Feng Zhou, Fernando Torre
Alignment of time series is an important problem to solve in many scientific dis ciplines. In particular, temporal alignment of two or more subjects performing s imilar activities is a challenging problem due to the large temporal scale diffe rence between human actions as well as the inter/intra subject variability. In t his paper we present canonical time warping (CTW), an extension of canonical cor relation analysis (CCA) for spatio-temporal alignment of the behavior between tw o subjects. CTW extends previous work on CCA in two ways: (i) it combines CCA wi th dynamic time warping for temporal alignment; and (ii) it extends CCA to allow local spatial deformations. We show CTWs effectiveness in three experiments: al ignment of synthetic data, alignment of motion capture data of two subjects perf orming similar actions, and alignment of two people with similar facial expressi ons. Our results demonstrate that CTW provides both visually and qualitatively b etter alignment than state-of-the-art techniques based on dynamic time warping.
****

A Parameter-free Hedging Algorithm
Kamalika Chaudhuri, Yoav Freund, Daniel J. Hsu
We study the problem of decision-theoretic online learning (DTOL). Motivated by practical applications, we focus on DTOL when the number of actions is very larg e. Previous algorithms for learning in this framework have a tunable learning ra te parameter, and a major barrier to using online-learning in practical applicat ions is that it is not understood how to set this parameter optimally, particula rly when the number of actions is large. In this paper, we offer a clean solutio n by proposing a novel and completely parameter-free algorithm for DTOL. In ad dition, we introduce a new notion of regret, which is more natural for applicati ons with a large number of actions. We show that our algorithm achieves good per formance with respect to this new notion of regret; in addition, it also achieve s performance close to that of the best bounds achieved by previous algorithms w ith optimally-tuned parameters, according to previous notions of regret.
****

Learning Bregman Distance Functions and Its Application for Semi-Supervised Clus tering
Lei Wu, Rong Jin, Steven Hoi, Jianke Zhu, Nenghai Yu
Learning distance functions with side information plays a key role in many machi ne learning and data mining applications. Conventional approaches often assume a Mahalanobis distance function. These approaches are limited in two aspects: (i) they are computationally expensive (even infeasible) for high dimensional data because the size of the metric is in the square of dimensionality; (ii) they ass ume a fixed metric for the entire input space and therefore are unable to handle heterogeneous data. In this paper, we propose a novel scheme that learns nonlin ear Bregman distance functions from side information using a non-parametric appr oach that is similar to support vector machines. The proposed scheme avoids the assumption of fixed metric because its local distance metric is implicitly deriv ed from the Hessian matrix of a convex function that is used to generate the Bre gman distance function. We present an efficient learning algorithm for the propo sed scheme for distance function learning. The extensive experiments with semi-s upervised clustering show the proposed technique (i) outperforms the state-of-th e-art approaches for distance function learning, and (ii) is computationally eff icient for high dimensional data.
****

Ranking Measures and Loss Functions in Learning to Rank
Wei Chen, Tie-yan Liu, Yanyan Lan, Zhi-ming Ma, Hang Li
Learning to rank has become an important research topic in machine learning. Whi
le most learning-to-rank methods learn the ranking function by minimizing the lo
ss functions, it is the ranking measures (such as NDCG and MAP) that are used to
 evaluate the performance of the learned ranking function. In this work, we reve
al the relationship between ranking measures and loss functions in learning-to-r
ank methods, such as Ranking SVM, RankBoost, RankNet, and ListMLE. We show that
these loss functions are upper bounds of the measure-based ranking errors. As a
result, the minimization of these loss functions will lead to the maximization o
f the ranking measures. The key to obtaining this result is to model ranking as
a sequence of classification tasks, and define a so-called essential loss as the
 weighted sum of the classification errors of individual tasks in the sequence.
We have proved that the essential loss is both an upper bound of the measure-bas
ed ranking errors, and a lower bound of the loss functions in the aforementioned
 methods. Our proof technique also suggests a way to modify existing loss functi
ons to make them tighter bounds of the measure-based ranking errors. Experimenta
l results on benchmark datasets show that the modifications can lead to better r
anking performance, demonstrating the correctness of our analysis.
****

Constructing Topological Maps using Markov Random Fields and Loop-Closure Detect
ion
Roy Anati, Kostas Daniilidis
We present a system which constructs a topological map of an environment given a
 sequence of images. This system includes a novel image similarity score which u
ses dynamic programming to match images using both the appearance and relative p
ositions of local features simultaneously. Additionally an MRF is constructed to
 model the probability of loop-closures. A locally optimal labeling is found usi
ng Loopy-BP. Finally we outline a method to generate a topological map from loop
 closure data. Results are presented on four urban sequences and one indoor sequ
ence.
****

Replicated Softmax: an Undirected Topic Model
Geoffrey E. Hinton, Russ R. Salakhutdinov
We show how to model documents as bags of words using family of two-layer, undir
ected graphical models. Each member of the family has the same number of binary
hidden units but a different number of ``softmax visible units. All of the softm
ax units in all of the models in the family share the same weights to the binary
 hidden units. We describe efficient inference and learning procedures for such
a family. Each member of the family models the probability distribution of docum
ents of a specific length as a product of topic-specific distributions rather th
an as a mixture and this gives much better generalization than Latent Dirichlet
Allocation for modeling the log probabilities of held-out documents. The low-dim
ensional topic vectors learned by the undirected family are also much better tha
n LDA topic vectors for retrieving documents that are similar to a query documen
t. The learned topics are more general than those found by LDA because precision
 is achieved by intersecting many general topics rather than by selecting a sing
le precise topic to generate each word.
****

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

Bayesian Belief Polarization
Alan Jern, Kai-min Chang, Charles Kemp
Situations in which people with opposing prior beliefs observe the same evidence
 and then strengthen those existing beliefs are frequently offered as evidence o

f human irrationality. This phenomenon, termed belief polarization, is typically assumed to be non-normative. We demonstrate, however, that a variety of cases of belief polarization are consistent with a Bayesian approach to belief revision. Simulation results indicate that belief polarization is not only possible but relatively common within the class of Bayesian models that we consider.
****

Linearly constrained Bayesian matrix factorization for blind source separation
Mikkel Schmidt
We present a general Bayesian approach to probabilistic matrix factorization subject to linear constraints. The approach is based on a Gaussian observation model and Gaussian priors with bilinear equality and inequality constraints. We present an efficient Markov chain Monte Carlo inference procedure based on Gibbs sampling. Special cases of the proposed model are Bayesian formulations of non-negative matrix factorization and factor analysis. The method is evaluated on a blind source separation problem. We demonstrate that our algorithm can be used to extract meaningful and interpretable features that are remarkably different from features extracted using existing related matrix factorization techniques.
****

Sparse and Locally Constant Gaussian Graphical Models
Jean Honorio, Dimitris Samaras, Nikos Paragios, Rita Goldstein, Luis E. Ortiz
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.
****

Sparse Metric Learning via Smooth Optimization
Yiming Ying, Kaizhu Huang, Colin Campbell
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.
****

The 'tree-dependent components' of natural scenes are edge filters
Daniel Zoran, Yair Weiss
We propose a new model for natural image statistics. Instead of minimizing dependency between components of natural images, we maximize a simple form of dependency in the form of tree-dependency. By learning filters and tree structures which are best suited for natural images we observe that the resulting filters are edge filters, similar to the famous ICA on natural images results. Calculating the likelihood of the model requires estimating the squared output of pairs of filters connected in the tree. We observe that after learning, these pairs of filters are predominantly of similar orientations but different phases, so their joint energy resembles models of complex cells.
****

Bootstrapping from Game Tree Search
Joel Veness, David Silver, Alan Blair, William Uther
In this paper we introduce a new algorithm for updating the parameters of a heuristic evaluation function, by updating the heuristic towards the values computed by an alpha-beta search. Our algorithm differs from previous approaches to learning from search, such as Samuels checkers player and the TD-Leaf algorithm, in two key ways. First, we update all nodes in the search tree, rather than a single node. Second, we use the outcome of a deep search, instead of the outcome of a subsequent search, as the training signal for the evaluation function. We implemented our algorithm in a chess program Meep, using a linear heuristic function. After initialising its weight vector to small random values, Meep was able to learn high quality weights from self-play alone. When tested online against human opponents, Meep played at a master level, the best performance of any chess program with a heuristic learned entirely from self-play.
****

Individuation, Identification and Object Discovery

Charles Kemp, Alan Jern, Fei Xu
Humans are typically able to infer how many objects their environment contains a nd to recognize when the same object is encountered twice. We present a simple statistical model that helps to explain these abilities and evaluate it in three behavioral experiments. Our first experiment suggests that humans rely on prio r knowledge when deciding whether an object token has been previously encountere d. Our second and third experiments suggest that humans can infer how many objec ts they have seen and can learn about categories and their properties even when they are uncertain about which tokens are instances of the same object.
****

Modeling Social Annotation Data with Content Relevance using a Topic Model
Tomoharu Iwata, Takeshi Yamada, Naonori Ueda
We propose a probabilistic topic model for analyzing and extracting content-rela ted annotations from noisy annotated discrete data such as web pages stored in s ocial bookmarking services. In these services, since users can attach annotation s freely, some annotations do not describe the semantics of the content, thus th ey are noisy, i.e. not content-related. The extraction of content-related anno tations can be used as a preprocessing step in machine learning tasks such as te xt classification and image recognition, or can improve information retrieval pe rformance. The proposed model is a generative model for content and annotations, in which the annotations are assumed to originate either from topics that gener ated the content or from a general distribution unrelated to the content. We dem onstrate the effectiveness of the proposed method by using synthetic data and re al social annotation data for text and images.
****

Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approxima tion
Hamid Maei, Csaba Szepesvári, Shalabh Bhatnagar, Doina Precup, David Silver, Ric hard S. Sutton
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Group Sparse Coding
Samy Bengio, Fernando Pereira, Yoram Singer, Dennis Strelow
Bag-of-words document representations are often used in text, image and video pr ocessing. While it is relatively easy to determine a suitable word dictionary fo r text documents, there is no simple mapping from raw images or videos to dictio nary terms. The classical approach builds a dictionary using vector quantization over a large set of useful visual descriptors extracted from a training set, an d uses a nearest-neighbor algorithm to count the number of occurrences of each d ictionary word in documents to be encoded. More robust approaches have been prop osed recently that represent each visual descriptor as a sparse weighted combina tion of dictionary words. While favoring a sparse representation at the level of visual descriptors, those methods however do not ensure that images have sparse representation. In this work, we use mixed-norm regularization to achieve spars ity at the image level as well as a small overall dictionary. This approach can also be used to encourage using the same dictionary words for all the images in a class, providing a discriminative signal in the construction of image represen tations. Experimental results on a benchmark image classification dataset show t hat when compact image or dictionary representations are needed for computationa l efficiency, the proposed approach yields better mean average precision in clas sification.
****

Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Proces s
Chong Wang, David Blei
We present a nonparametric hierarchical Bayesian model of document collections t hat decouples sparsity and smoothness in the component distributions (i.e., the

``topics). In the sparse topic model (STM), each topic is represented by a bank of selector variables that determine which terms appear in the topic. Thus each topic is associated with a subset of the vocabulary, and topic smoothness is mod eled on this subset. We develop an efficient Gibbs sampler for the STM that incl udes a general-purpose method for sampling from a Dirichlet mixture with a combi natorial number of components. We demonstrate the STM on four real-world dataset s. Compared to traditional approaches, the empirical results show that STMs give better predictive performance with simpler inferred models.
****

Fast Image Deconvolution using Hyper-Laplacian Priors
Dilip Krishnan, Rob Fergus
The heavy-tailed distribution of gradients in natural scenes have proven effecti ve priors for a range of problems such as denoising, deblurring and super-resolu tion. However, the use of sparse distributions makes the problem non-convex and impractically slow to solve for multi-megapixel images. In this paper we describ e a deconvolution approach that is several orders of magnitude faster than exist ing techniques that use hyper-Laplacian priors. We adopt an alternating minimiza tion scheme where one of the two phases is a non-convex problem that is separabl e over pixels. This per-pixel sub-problem may be solved with a lookup table (LUT). Alternatively, for two specific values of α, 1/2 and 2/3 an analytic solution can be found, by finding the roots of a cubic and quartic polynomial, respectiv ely. Our approach (using either LUTs or analytic formulae) is able to deconvolve a 1 megapixel image in less than ~3 seconds, achieving comparable quality to ex isting methods such as iteratively reweighted least squares (IRLS) that take ~20 minutes. Furthermore, our method is quite general and can easily be extended to related image processing problems, beyond the deconvolution application demonst rated.
****

Compositionality of optimal control laws
Emanuel Todorov
We present a theory of compositionality in stochastic optimal control, showing h ow task-optimal controllers can be constructed from certain primitives. The prim itives are themselves feedback controllers pursuing their own agendas. They are mixed in proportion to how much progress they are making towards their agendas a nd how compatible their agendas are with the present task. The resulting composi te control law is provably optimal when the problem belongs to a certain class. This class is rather general and yet has a number of unique properties - one of which is that the Bellman equation can be made linear even for non-linear or dis crete dynamics. This gives rise to the compositionality developed here. In the s pecial case of linear dynamics and Gaussian noise our framework yields analytica l solutions (i.e. non-linear mixtures of linear-quadratic regulators) without re quiring the final cost to be quadratic. More generally, a natural set of control primitives can be constructed by applying SVD to Greens function of the Bellman equation. We illustrate the theory in the context of human arm movements. The i deas of optimality and compositionality are both very prominent in the field of motor control, yet they are hard to reconcile. Our work makes this possible.
****

AUC optimization and the two-sample problem
Nicolas Vayatis, Marine Depecker, Stéphan Clémençcon
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Measuring Invariances in Deep Networks
Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, Andrew Ng
For many computer vision applications, the ideal image feature would be invarian t to multiple confounding image properties, such as illumination and viewing ang le. Recently, deep architectures trained in an unsupervised manner have been pro posed as an automatic method for extracting useful features. However, outside of

using these learning algorithms in a classi■er, they can be sometimes dif■cult to evaluate. In this paper, we propose a number of empirical tests that directly measure the degree to which these learned features are invariant to different image transforms. We ■nd that deep autoencoders become invariant to increasingly complex image transformations with depth. This further justi■es the use of “deep” vs. “shallower” representations. Our performance metrics agree with existing measures of invariance. Our evaluation metrics can also be used to evaluate future work in unsupervised deep learning, and thus help the development of future algorithms.

****

Nonparametric Latent Feature Models for Link Prediction

Kurt Miller, Michael Jordan, Thomas Griffiths

As the availability and importance of relational data -- such as the friendships summarized on a social networking website -- increases, it becomes increasingly important to have good models for such data. The kinds of latent structure that have been considered for use in predicting links in such networks have been relatively limited. In particular, the machine learning community has focused on latent class models, adapting nonparametric Bayesian methods to jointly infer how many latent classes there are while learning which entities belong to each class. We pursue a similar approach with a richer kind of latent variable -- latent features -- using a nonparametric Bayesian technique to simultaneously infer the number of features at the same time we learn which entities have each feature. The greater expressiveness of this approach allows us to improve link prediction on three datasets.

****

Asymptotically Optimal Regularization in Smooth Parametric Models

Percy S. Liang, Guillaume Bouchard, Francis Bach, Michael Jordan

Many types of regularization schemes have been employed in statistical learning, each one motivated by some assumption about the problem domain. In this paper, we present a unified asymptotic analysis of smooth regularizers, which allows us to see how the validity of these assumptions impacts the success of a particular regularizer. In addition, our analysis motivates an algorithm for optimizing regularization parameters, which in turn can be analyzed within our framework. We apply our analysis to several examples, including hybrid generative-discriminative learning and multi-task learning.

****

Variational Gaussian-process factor analysis for modeling spatio-temporal data

Jaakko Luttinen, Alexander Ilin

We present a probabilistic latent factor model which can be used for studying spatio-temporal datasets. The spatial and temporal structure is modeled by using Gaussian process priors both for the loading matrix and the factors. The posterior distributions are approximated using the variational Bayesian framework. High computational cost of Gaussian process modeling is reduced by using sparse approximations. The model is used to compute the reconstructions of the global sea surface temperatures from a historical dataset. The results suggest that the proposed model can outperform the state-of-the-art reconstruction systems.

****

Lattice Regression

Eric Garcia, Maya Gupta

We present a new empirical risk minimization framework for approximating functions from training samples for low-dimensional regression applications where a lattice (look-up table) is stored and interpolated at run-time for an efficient hardware implementation. Rather than evaluating a fitted function at the lattice nodes without regard to the fact that samples will be interpolated, the proposed lattice regression approach estimates the lattice to minimize the interpolation error on the given training samples. Experiments show that lattice regression can reduce mean test error compared to Gaussian process regression for digital color management of printers, an application for which linearly interpolating a look-up table (LUT) is standard. Simulations confirm that lattice regression performs consistently better than the naive approach to learning the lattice, particul

arly when the density of training samples is low.
****
Sharing Features among Dynamical Systems with Beta Processes
Emily Fox, Michael Jordan, Erik Sudderth, Alan Willsky
We propose a Bayesian nonparametric approach to relating multiple time series via a set of latent, dynamical behaviors. Using a beta process prior, we allow data-driven selection of the size of this set, as well as the pattern with which behaviors are shared among time series. Via the Indian buffet process representation of the beta process predictive distributions, we develop an exact Markov chain Monte Carlo inference method. In particular, our approach uses the sum-product algorithm to efficiently compute Metropolis-Hastings acceptance probabilities, and explores new dynamical behaviors via birth/death proposals. We validate our sampling algorithm using several synthetic datasets, and also demonstrate promising unsupervised segmentation of visual motion capture data.
****
The Wisdom of Crowds in the Recollection of Order Information
Mark Steyvers, Brent Miller, Pernille Hemmer, Michael Lee
When individuals independently recollect events or retrieve facts from memory, how can we aggregate these retrieved memories to reconstruct the actual set of events or facts? In this research, we report the performance of individuals in a series of general knowledge tasks, where the goal is to reconstruct from memory the order of historic events, or the order of items along some physical dimension. We introduce two Bayesian models for aggregating order information based on a Thurstonian approach and Mallows model. Both models assume that each individuals reconstruction is based on either a random permutation of the unobserved ground truth, or by a pure guessing strategy. We apply MCMC to make inferences about the underlying truth and the strategies employed by individuals. The models demonstrate a wisdom of crowds" effect, where the aggregated orderings are closer to the true ordering than the orderings of the best individual."
****
Efficient Recovery of Jointly Sparse Vectors
Liang Sun, Jun Liu, Jianhui Chen, Jieping Ye
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.
****
Efficient Match Kernel between Sets of Features for Visual Recognition
Liefeng Bo, Cristian Sminchisescu
In visual recognition, the images are frequently modeled as sets of local features (bags). We show that bag of words, a common method to handle such cases, can be viewed as a special match kernel, which counts 1 if two local features fall into the same regions partitioned by visual words and 0 otherwise. Despite its simplicity, this quantization is too coarse. It is, therefore, appealing to design match kernels that more accurately measure the similarity between local features. However, it is impractical to use such kernels on large datasets due to their significant computational cost. To address this problem, we propose an efficient match kernel (EMK), which maps local features to a low dimensional feature space, average the resulting feature vectors to form a set-level feature, then apply a linear classifier. The local feature maps are learned so that their inner products preserve, to the best possible, the values of the specified kernel function. EMK is linear both in the number of images and in the number of local features. We demonstrate that EMK is extremely efficient and achieves the current state of the art performance on three difficult real world datasets: Scene-15, Caltech-101 and Caltech-256.
****
Clustering sequence sets for motif discovery
Jong Kim, Seungjin Choi
Most of existing methods for DNA motif discovery consider only a single set of sequences to find an over-represented motif. In contrast, we consider multiple se

ts of sequences where we group sets associated with the same motif into a cluste
r, assuming that each set involves a single motif. Clustering sets of sequences
yields clusters of coherent motifs, improving signal-to-noise ratio or enabling
us to identify multiple motifs. We present a probabilistic model for DNA motif d
iscovery where we identify multiple motifs through searching for patterns which
are shared across multiple sets of sequences. Our model infers cluster-indicatin
g latent variables and learns motifs simultaneously, where these two tasks inter
act with each other. We show that our model can handle various motif discovery p
roblems, depending on how to construct multiple sets of sequences. Experiments o
n three different problems for discovering DNA motifs emphasize the useful behav
ior and confirm the substantial gains over existing methods where only single se
t of sequences is considered.
****

Bayesian Nonparametric Models on Decomposable Graphs
Francois Caron, Arnaud Doucet
Over recent years Dirichlet processes and the associated Chinese restaurant proc
ess (CRP) have found many applications in clustering while the Indian buffet pro
cess (IBP) is increasingly used to describe latent feature models. In the cluste
ring case, we associate to each data point a latent allocation variable. These l
atent variables can share the same value and this induces a partition of the dat
a set. The CRP is a prior distribution on such partitions. In latent feature mo
dels, we associate to each data point a potentially infinite number of binary la
tent variables indicating the possession of some features and the IBP is a prior
 distribution on the associated infinite binary matrix. These prior distribution
s are attractive because they ensure exchangeability (over samples). We propose
here extensions of these models to decomposable graphs. These models have appeal
ing properties and can be easily learned using Monte Carlo techniques.
****

Correlation Coefficients are Insufficient for Analyzing Spike Count Dependencies
Arno Onken, Steffen Grünewälder, Klaus Obermayer
The linear correlation coefficient is typically used to characterize and analyze
 dependencies of neural spike counts. Here, we show that the correlation coeffic
ient is in general insufficient to characterize these dependencies. We construct
 two neuron spike count models with Poisson-like marginals and vary their depend
ence structure using copulas. To this end, we construct a copula that allows to
keep the spike counts uncorrelated while varying their dependence strength. More
over, we employ a network of leaky integrate-and-fire neurons to investigate whe
ther weakly correlated spike counts with strong dependencies are likely to occur
 in real networks. We find that the entropy of uncorrelated but dependent spike
count distributions can deviate from the corresponding distribution with indepen
dent components by more than 25% and that weakly correlated but strongly depende
nt spike counts are very likely to occur in biological networks. Finally, we int
roduce a test for deciding whether the dependence structure of distributions wit
h Poisson-like marginals is well characterized by the linear correlation coeffic
ient and verify it for different copula-based models.
****

Streaming k-means approximation
Nir Ailon, Ragesh Jaiswal, Claire Monteleoni
We provide a clustering algorithm that approximately optimizes the k-means objec
tive, in the one-pass streaming setting. We make no assumptions about the data,
and our algorithm is very light-weight in terms of memory, and computation. Thi
s setting is applicable to unsupervised learning on massive data sets, or resour
ce-constrained devices. The two main ingredients of our theoretical work are: a
 derivation of an extremely simple pseudo-approximation batch algorithm for k-me
ans, in which the algorithm is allowed to output more than k centers (based on t
he recent k-means++"), and a streaming clustering algorithm in which batch clust
ering algorithms are performed on small inputs (fitting in memory) and combined
in a hierarchical manner. Empirical evaluations on real and simulated data reve
al the practical utility of our method."
****

Help or Hinder: Bayesian Models of Social Goal Inference
Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, Joshua Tenenbaum
Everyday social interactions are heavily influenced by our snap judgments about others goals. Even young infants can infer the goals of intentional agents from observing how they interact with objects and other agents in their environment: e.g., that one agent is helping orhindering anothers attempt to get up a hill or open a box. We propose a model for how people can infer these social goals from actions, based on inverse planning in multiagent Markov decision problems (MDPs). The model infers the goal most likely to be driving an agents behavior by assuming the agent acts approximately rationally given environmental constraints and its model of other agents present. We also present behavioral evidence in support of this model over a simpler, perceptual cue-based alternative.
****
Sequential effects reflect parallel learning of multiple environmental regularities
Matthew Wilder, Matt Jones, Michael C. Mozer
Across a wide range of cognitive tasks, recent experience in■uences behavior. For example, when individuals repeatedly perform a simple two-alternative forced-choice task (2AFC), response latencies vary dramatically based on the immediately preceding trial sequence. These sequential effects have been interpreted as adaptation to the statistical structure of an uncertain, changing environment (e.g. Jones & Sieck, 2003; Mozer, Kinoshita, & Shettel, 2007; Yu & Cohen, 2008). The Dynamic Belief Model (DBM) (Yu & Cohen, 2008) explains sequential effects in 2AFC tasks as a rational consequence of a dynamic internal representation that tracks second-order statistics of the trial sequence (repetition rates) and predicts whether the upcoming trial will be a repetition or an alternation of the previous trial. Experimental results suggest that ■rst-order statistics (base rates) also in■uence sequential effects. We propose a model that learns both ■rst- and second-order sequence properties, each according to the basic principles of the DBM but under a uni■ed inferential framework. This model, the Dynamic Belief Mixture Model (DBM2), obtains precise, parsimonious ■ts to data. Furthermore, the model predicts dissociations in behavioral (Maloney, Dal Martello, Sahm, & Spillmann, 2005) and electrophysiological studies (Jentzsch & Sommer, 2002), supporting the psychological and neurobiological reality of its two components.
****
Noisy Generalized Binary Search
Robert Nowak
This paper addresses the problem of noisy Generalized Binary Search (GBS). GBS is a well-known greedy algorithm for determining a binary-valued hypothesis through a sequence of strategically selected queries. At each step, a query is selected that most evenly splits the hypotheses under consideration into two disjoint subsets, a natural generalization of the idea underlying classic binary search. GBS is used in many applications, including fault testing, machine diagnostics, disease diagnosis, job scheduling, image processing, computer vision, and active learning. In most of these cases, the responses to queries can be noisy. Past work has provided a partial characterization of GBS, but existing noise-tolerant versions of GBS are suboptimal in terms of sample complexity. This paper presents the first optimal algorithm for noisy GBS and demonstrates its application to learning multidimensional threshold functions.
****
Solving Stochastic Games
Liam Dermed, Charles Isbell
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.
****
Modelling Relational Data using Bayesian Clustered Tensor Factorization
Ilya Sutskever, Joshua Tenenbaum, Russ R. Salakhutdinov

We consider the problem of learning probabilistic models for complex relational structures between various types of objects. A model can help us ``understand a dataset of relational facts in at least two ways, by finding interpretable structure in the data, and by supporting predictions, or inferences about whether particular unobserved relations are likely to be true. Often there is a tradeoff between these two aims: cluster-based models yield more easily interpretable representations, while factorization-based approaches have better predictive performance on large data sets. We introduce the Bayesian Clustered Tensor Factorization (BCTF) model, which embeds a factorized representation of relations in a nonparametric Bayesian clustering framework. Inference is fully Bayesian but scales well to large data sets. The model simultaneously discovers interpretable clusters and yields predictive performance that matches or beats previous probabilistic models for relational data.
****

Kernel Methods for Deep Learning
Youngmin Cho, Lawrence Saul
We introduce a new family of positive-definite kernel functions that mimic the computation in large, multilayer neural nets. These kernel functions can be used in shallow architectures, such as support vector machines (SVMs), or in deep kernel-based architectures that we call multilayer kernel machines (MKMs). We evaluate SVMs and MKMs with these kernel functions on problems designed to illustrate the advantages of deep architectures. On several problems, we obtain better results than previous, leading benchmarks from both SVMs with Gaussian kernels as well as deep belief nets.
****

Optimal context separation of spiking haptic signals by second-order somatosensory neurons
Romain Brasselet, Roland Johansson, Angelo Arleo
We study an encoding/decoding mechanism accounting for the relative spike timing of the signals propagating from peripheral nerve fibers to second-order somatosensory neurons in the cuneate nucleus (CN). The CN is modeled as a population of spiking neurons receiving as inputs the spatiotemporal responses of real mechanoreceptors obtained via microneurography recordings in humans. The efficiency of the haptic discrimination process is quantified by a novel definition of entropy that takes into full account the metrical properties of the spike train space. This measure proves to be a suitable decoding scheme for generalizing the classical Shannon entropy to spike-based neural codes. It permits an assessment of neurotransmission in the presence of a large output space (i.e. hundreds of spike trains) with 1 ms temporal precision. It is shown that the CN population code performs a complete discrimination of 81 distinct stimuli already within 35 ms of the first afferent spike, whereas a partial discrimination (80% of the maximum information transmission) is possible as rapidly as 15 ms. This study suggests that the CN may not constitute a mere synaptic relay along the somatosensory pathway but, rather, it may convey optimal contextual accounts (in terms of fast and reliable information transfer) of peripheral tactile inputs to downstream structures of the central nervous system.
****

Heterogeneous multitask learning with joint sparsity constraints
Xiaolin Yang, Seyoung Kim, Eric Xing
Multitask learning addressed the problem of learning related tasks whose information can be shared each other. Traditional problem usually deal with homogeneous tasks such as regression, classification individually. In this paper we consider the problem learning multiple related tasks where tasks consist of both continuous and discrete outputs from a common set of input variables that lie in a high-dimensional space. All of the tasks are related in the sense that they share the same set of relevant input variables, but the amount of influence of each input on different outputs may vary. We formulate this problem as a combination of linear regression and logistic regression and model the joint sparsity as L1/Linf and L1/L2-norm of the model parameters. Among several possible applications, our approach addresses an important open problem in genetic association mapping,

where we are interested in discovering genetic markers that influence multiple c
orrelated traits jointly. In our experiments, we demonstrate our method in the s
cenario of association mapping, using simulated and asthma data, and show that t
he algorithm can effectively recover the relevant inputs with respect to all of
the tasks.
****

Learning with Compressible Priors
Volkan Cevher
We describe probability distributions, dubbed compressible priors, whose indepen
dent and identically distributed (iid) realizations result in compressible signa
ls. A signal is compressible when sorted magnitudes of its coefficients exhibit
a power-law decay so that the signal can be well-approximated by a sparse signal
. Since compressible signals live close to sparse signals, their intrinsic infor
mation can be stably embedded via simple non-adaptive linear projections into a
much lower dimensional space whose dimension grows logarithmically with the ambi
ent signal dimension. By using order statistics, we show that N-sample iid reali
zations of generalized Pareto, Student’s t, log-normal, Frechet, and log-logisti
c distributions are compressible, i.e., they have a constant expected decay rate
, which is independent of N. In contrast, we show that generalized Gaussian dist
ribution with shape parameter q is compressible only in restricted cases since t
he expected decay rate of its N-sample iid realizations decreases with N as 1/[q
 log(N/q)]. We use compressible priors as a scaffold to build new iterative spar
se signal recovery algorithms based on Bayesian inference arguments. We show how
 tuning of these algorithms explicitly depends on the parameters of the compress
ible prior of the signal, and how to learn the parameters of the signal’s compre
ssible prior on the fly during recovery.
****

Inter-domain Gaussian Processes for Sparse Inference using Inducing Features
Miguel Lázaro-Gredilla, Aníbal Figueiras-Vidal
We present a general inference framework for inter-domain Gaussian Processes (GP
s), focusing on its usefulness to build sparse GP models. The state-of-the-art s
parse GP model introduced by Snelson and Ghahramani in [1] relies on finding a s
mall, representative pseudo data set of m elements (from the same domain as the
n available data elements) which is able to explain existing data well, and then
 uses it to perform inference. This reduces inference and model selection comput
ation time from O(n^3) to O(m^2n), where m << n. Inter-domain GPs can be used to
 find a (possibly more compact) representative set of features lying in a differ
ent domain, at the same computational cost. Being able to specify a different do
main for the representative features allows to incorporate prior knowledge about
 relevant characteristics of data and detaches the functional form of the covari
ance and basis functions. We will show how previously existing models fit into t
his framework and will use it to develop two new sparse GP models. Tests on larg
e, representative regression data sets suggest that significant improvement can
be achieved, while retaining computational efficiency.
****

Particle-based Variational Inference for Continuous Systems
Andrew Frank, Padhraic Smyth, Alexander Ihler
Since the development of loopy belief propagation, there has been considerable w
ork on advancing the state of the art for approximate inference over distributio
ns defined on discrete random variables. Improvements include guarantees of conv
ergence, approximations that are provably more accurate, and bounds on the resul
ts of exact inference. However, extending these methods to continuous-valued s
ystems has lagged behind. While several methods have been developed to use beli
ef propagation on systems with continuous values, they have not as yet incorpora
ted the recent advances for discrete variables. In this context we extend a rece
ntly proposed particle-based belief propagation algorithm to provide a general f
ramework for adapting discrete message-passing algorithms to perform inference i
n continuous systems. The resulting algorithms behave similarly to their purely
 discrete counterparts, extending the benefits of these more advanced inference
techniques to the continuous domain.

****
Efficient Learning using Forward-Backward Splitting
Yoram Singer, John C. Duchi
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.
****
L_1-Penalized Robust Estimation for a Class of Inverse Problems Arising in Multiview Geometry
Arnak Dalalyan, Renaud Keriven
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.
****
Learning to Hash with Binary Reconstructive Embeddings
Brian Kulis, Trevor Darrell
Fast retrieval methods are increasingly critical for many large-scale analysis tasks, and there have been several recent methods that attempt to learn hash functions for fast and accurate nearest neighbor searches. In this paper, we develop an algorithm for learning hash functions based on explicitly minimizing the reconstruction error between the original distances and the Hamming distances of the corresponding binary embeddings. We develop a scalable coordinate-descent algorithm for our proposed hashing objective that is able to efficiently learn hash functions in a variety of settings. Unlike existing methods such as semantic hashing and spectral hashing, our method is easily kernelized and does not require restrictive assumptions about the underlying distribution of the data. We present results over several domains to demonstrate that our method outperforms existing state-of-the-art techniques.
****
Multi-Label Prediction via Compressed Sensing
Daniel J. Hsu, Sham M. Kakade, John Langford, Tong Zhang
We consider multi-label prediction problems with large output spaces under the assumption of output sparsity – that the target (label) vectors have small support. We develop a general theory for a variant of the popular error correcting output code scheme, using ideas from compressed sensing for exploiting this sparsity. The method can be regarded as a simple reduction from multi-label regression problems to binary regression problems. We show that the number of subprob- lems need only be logarithmic in the total number of possible labels, making this approach radically more ef■cient than others. We also state and prove robustness guarantees for this method in the form of regret transform bounds (in general), and also provide a more detailed analysis for the linear prediction setting.
****
Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions
Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, Bharath K. Sriperumbudur
Embeddings of probability measures into reproducing kernel Hilbert spaces have been proposed as a straightforward and practical means of representing and comparing probabilities. In particular, the distance between embeddings (the maximum mean discrepancy, or MMD) has several key advantages over many classical metrics on distributions, namely easy computability, fast convergence and low bias of finite sample estimates. An important requirement of the embedding RKHS is that it be characteristic: in this case, the MMD between two distributions is zero if and only if the distributions coincide. Three new results on the MMD are introduced in the present study. First, it is established that MMD corresponds to the optimal risk of a kernel classifier, thus forming a natural link between the distance between distributions and their ease of classification. An important consequence is that a kernel must be characteristic to guarantee classifiability betwee

n distributions in the RKHS. Second, the class of characteristic kernels is broa
dened to incorporate all strictly positive definite kernels: these include non-t
ranslation invariant kernels and kernels on non-compact domains. Third, a genera
lization of the MMD is proposed for families of kernels, as the supremum over MM
Ds on a class of kernels (for instance the Gaussian kernels with different bandw
idths). This extension is necessary to obtain a single distance measure if a lar
ge selection or class of characteristic kernels is potentially appropriate. This
 generalization is reasonable, given that it corresponds to the problem of learn
ing the kernel by minimizing the risk of the corresponding kernel classifier. Th
e generalized MMD is shown to have consistent finite sample estimates, and its p
erformance is demonstrated on a homogeneity testing example.
****

Statistical Analysis of Semi-Supervised Learning: The Limit of Infinite Unlabell
ed Data

Boaz Nadler, Nathan Srebro, Xueyuan Zhou

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

Maximin affinity learning of image segmentation

Kevin Briggman, Winfried Denk, Sebastian Seung, Moritz Helmstaedter, Srinivas C.
 Turaga

Images can be segmented by first using a classifier to predict an affinity graph
 that reflects the degree to which image pixels must be grouped together and the
n partitioning the graph to yield a segmentation. Machine learning has been appl
ied to the affinity classifier to produce affinity graphs that are good in the s
ense of minimizing edge misclassification rates. However, this error measure is
 only indirectly related to the quality of segmentations produced by ultimately p
artitioning the affinity graph. We present the first machine learning algorithm
 for training a classifier to produce affinity graphs that are good in the sense
 of producing segmentations that directly minimize the Rand index, a well known s
egmentation performance measure. The Rand index measures segmentation performanc
e by quantifying the classification of the connectivity of image pixel pairs aft
er segmentation. By using the simple graph partitioning algorithm of finding the
 connected components of the thresholded affinity graph, we are able to train an
 affinity classifier to directly minimize the Rand index of segmentations result
ing from the graph partitioning. Our learning algorithm corresponds to the learn
ing of maximin affinities between image pixel pairs, which are predictive of the
 pixel-pair connectivity.
****

Perceptual Multistability as Markov Chain Monte Carlo Inference

Samuel Gershman, Ed Vul, Joshua Tenenbaum

While many perceptual and cognitive phenomena are well described in terms of Bay
esian inference, the necessary computations are intractable at the scale of real
-world tasks, and it remains unclear how the human mind approximates Bayesian in
ference algorithmically. We explore the proposal that for some tasks, humans use
 a form of Markov Chain Monte Carlo to approximate the posterior distribution ov
er hidden variables. As a case study, we show how several phenomena of perceptu
al multistability can be explained as MCMC inference in simple graphical models
 for low-level vision.
****

Probabilistic Relational PCA

Wu-jun Li, Dit-Yan Yeung, Zhihua Zhang

One crucial assumption made by both principal component analysis (PCA) and proba
bilistic PCA (PPCA) is that the instances are independent and identically distri
buted (i.i.d.). However, this common i.i.d. assumption is unreasonable for relat
ional data. In this paper, by explicitly modeling covariance between instances a
s derived from the relational information, we propose a novel probabilistic dime
nsionality reduction method, called probabilistic relational PCA (PRPCA), for re

lational data analysis. Although the i.i.d. assumption is no longer adopted in P
RPCA, the learning algorithms for PRPCA can still be devised easily like those f
or PPCA which makes explicit use of the i.i.d. assumption. Experiments on real-w
orld data sets show that PRPCA can effectively utilize the relational informatio
n to dramatically outperform PCA and achieve state-of-the-art performance.
****

Adapting to the Shifting Intent of Search Queries
Umar Syed, Aleksandrs Slivkins, Nina Mishra
Search engines today present results that are often oblivious to recent shifts i
n intent. For example, the meaning of the query independence day shifts in early
 July to a US holiday and to a movie around the time of the box office release.
 While no studies exactly quantify the magnitude of intent-shifting traffic, stu
dies suggest that news events, seasonal topics, pop culture, etc account for 1/2
 the search queries. This paper shows that the signals a search engine receives
 can be used to both determine that a shift in intent happened, as well as find
a result that is now more relevant. We present a meta-algorithm that marries a c
lassifier with a bandit algorithm to achieve regret that depends logarithmically
 on the number of query impressions, under certain assumptions. We provide stro
ng evidence that this regret is close to the best achievable. Finally, via a ser
ies of experiments, we demonstrate that our algorithm outperforms prior approach
es, particularly as the amount of intent-shifting traffic increases.
****

Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory
Harold Pashler, Nicholas Cepeda, Robert V. Lindsey, Ed Vul, Michael C. Mozer
When individuals learn facts (e.g., foreign language vocabulary) over multiple s
tudy sessions, the temporal spacing of study has a significant impact on memory
retention. Behavioral experiments have shown a nonmonotonic relationship betwee
n spacing and retention: short or long intervals between study sessions yield l
ower cued-recall accuracy than intermediate intervals. Appropriate spacing of s
tudy can double retention on educationally relevant time scales. We introduce a
 Multiscale Context Model (MCM) that is able to predict the influence of a parti
cular study schedule on retention for specific material. MCMs prediction is bas
ed on empirical data characterizing forgetting of the material following a singl
e study session. MCM is a synthesis of two existing memory models (Staddon, Che
laru, & Higa, 2002; Raaijmakers, 2003). On the surface, these models are unrel
ated and incompatible, but we show they share a core feature that allows them t
o be integrated. MCM can determine study schedules that maximize the durabilit
y of learning, and has implications for education and training. MCM can be cas
t either as a neural network with inputs that fluctuate over time, or as a casc
ade of leaky integrators. MCM is intriguingly similar to a Bayesian multiscale
 model of memory (Kording, Tenenbaum, Shadmehr, 2007), yet MCM is better able to
 account for human declarative memory.
****

A Game-Theoretic Approach to Hypergraph Clustering
Samuel Bulò, Marcello Pelillo
Hypergraph clustering refers to the process of extracting maximally coherent gro
ups from a set of objects using high-order (rather than pairwise) similarities.
Traditional approaches to this problem are based on the idea of partitioning the
 input data into a user-defined number of classes, thereby obtaining the cluster
s as a by-product of the partitioning process. In this paper, we provide a radic
ally different perspective to the problem. In contrast to the classical approach
, we attempt to provide a meaningful formalization of the very notion of a clust
er and we show that game theory offers an attractive and unexplored perspective
that serves well our purpose. Specifically, we show that the hypergraph clusteri
ng problem can be naturally cast into a non-cooperative multi-player ``clusterin
g game, whereby the notion of a cluster is equivalent to a classical game-theore
tic equilibrium concept. From the computational viewpoint, we show that the prob
lem of finding the equilibria of our clustering game is equivalent to locally op
timizing a polynomial function over the standard simplex, and we provide a discr
ete-time dynamics to perform this optimization. Experiments are presented which

show the superiority of our approach over state-of-the-art hypergraph clustering techniques.

****

Dirichlet-Bernoulli Alignment: A Generative Model for Multi-Class Multi-Label Multi-Instance Corpora

Shuang-hong Yang, Hongyuan Zha, Bao-gang Hu

We propose Dirichlet-Bernoulli Alignment (DBA), a generative model for corpora in which each pattern (e.g., a document) contains a set of instances (e.g., paragraphs in the document) and belongs to multiple classes. By casting predefined classes as latent Dirichlet variables (i.e., instance level labels), and modeling the multi-label of each pattern as Bernoulli variables conditioned on the weighted empirical average of topic assignments, DBA automatically aligns the latent topics discovered from data to human-defined classes. DBA is useful for both pattern classification and instance disambiguation, which are tested on text classification and named entity disambiguation for web search queries respectively.

****

Adaptive Design Optimization in Experiments with People

Daniel Cavagnaro, Jay Myung, Mark Pitt

In cognitive science, empirical data collected from participants are the arbiters in model selection. Model discrimination thus depends on designing maximally informative experiments. It has been shown that adaptive design optimization (ADO) allows one to discriminate models as efficiently as possible in simulation experiments. In this paper we use ADO in a series of experiments with people to discriminate the Power, Exponential, and Hyperbolic models of memory retention, which has been a long-standing problem in cognitive science, providing an ideal setting in which to test the application of ADO for addressing questions about human cognition. Using an optimality criterion based on mutual information, ADO is able to find designs that are maximally likely to increase our certainty about the true model upon observation of the experiment outcomes. Results demonstrate the usefulness of ADO and also reveal some challenges in its implementation.

****

Riffled Independence for Ranked Data

Jonathan Huang, Carlos Guestrin

Representing distributions over permutations can be a daunting task due to the fact that the number of permutations of n objects scales factorially in n. One recent way that has been used to reduce storage complexity has been to exploit probabilistic independence, but as we argue, full independence assumptions impose strong sparsity constraints on distributions and are unsuitable for modeling rankings. We identify a novel class of independence structures, called riffled independence, which encompasses a more expressive family of distributions while retaining many of the properties necessary for performing efficient inference and reducing sample complexity. In riffled independence, one draws two permutations independently, then performs the riffle shuffle, common in card games, to combine the two permutations to form a single permutation. In ranking, riffled independence corresponds to ranking disjoint sets of objects independently, then interleaving those rankings. We provide a formal introduction and present algorithms for using riffled independence within Fourier-theoretic frameworks which have been explored by a number of recent papers.

****

A Neural Implementation of the Kalman Filter

Robert Wilson, Leif Finkel

There is a growing body of experimental evidence to suggest that the brain is capable of approximating optimal Bayesian inference in the face of noisy input stimuli. Despite this progress, the neural underpinnings of this computation are still poorly understood. In this paper we focus on the problem of Bayesian filtering of stochastic time series. In particular we introduce a novel neural network, derived from a line attractor architecture, whose dynamics map directly onto those of the Kalman Filter in the limit where the prediction error is small. When the prediction error is large we show that the network responds robustly to change-points in a way that is qualitatively compatible with the optimal Bayesi

an model. The model suggests ways in which probability distributions are encode d in the brain and makes a number of testable experimental predictions.
****
An LP View of the M-best MAP problem
Menachem Fromer, Amir Globerson
We consider the problem of finding the M assignments with maximum probability in a probabilistic graphical model. We show how this problem can be formulated as a linear program (LP) on a particular polytope. We prove that, for tree graphs (and junction trees in general), this polytope has a particularly simple form and differs from the marginal polytope in a single inequality constraint. We use th is characterization to provide an approximation scheme for non-tree graphs, by u sing the set of spanning trees over such graphs. The method we present puts the M -best inference problem in the context of LP relaxations, which have recently received considerable attention and have proven useful in solving difficult infe rence problems. We show empirically that our method often finds the provably exa ct M best configurations for problems of high tree-width. A common task in proba bilistic modeling is finding the assignment with maximum probability given a mod el. This is often referred to as the MAP (maximum a-posteriori) problem. Of part icular interest is the case of MAP in graphical models, i.e., models where the p robability factors into a product over small subsets of variables. For general m odels, this is an NP-hard problem [11], and thus approximation algorithms are re quired. Of those, the class of LP based relaxations has recently received consid erable attention [3, 5, 18]. In fact, it has been shown that some problems (e.g. , fixed backbone protein design) can be solved exactly via sequences of increasi ngly tighter LP relaxations [13]. In many applications, one is interested not on ly in the MAP assignment but also in the M maximum probability assignments [19]. For example, in a protein design problem, we might be interested in the M amino acid sequences that are most stable on a given backbone structure [2]. In cases where the MAP problem is tractable, one can devise tractable algorithms for the M best problem [8, 19]. Specifically, for low tree-width graphs, this can be do ne via a variant of max-product [19]. However, when finding MAPs is not tractabl e, it is much less clear how to approximate the M best case. One possible approa ch is to use loopy max-product to obtain approximate max-marginals and use those to approximate the M best solutions [19]. However, this is largely a heuristic and does not provide any guarantees in terms of optimality certificates or bound s on the optimal values. LP approximations to MAP do enjoy such guarantees. Spec ifically, they provide upper bounds on the MAP value and optimality certificates . Furthermore, they often work for graphs with large tree-width [13]. The goal o f the current work is to leverage the power of LP relaxations to the M best case . We begin by focusing on the problem of finding the second best solution. We sh ow how it can be formulated as an LP over a polytope we call the "assignment-exc luding marginal polytope". In the general case, this polytope may require an exp onential number of inequalities, but we prove that when the graph is a tree it h as a very compact representation. We proceed to use this result to obtain approx imations to the second best problem, and show how these can be tightened in vari ous ways. Next, we show how M best assignments can be found by relying on algori thms for 1
****
Speeding up Magnetic Resonance Image Acquisition by Bayesian Multi-Slice Adaptiv e Compressed Sensing
Matthias Seeger
We show how to sequentially optimize magnetic resonance imaging measurement desi gns over stacks of neighbouring image slices, by performing convex variational i nference on a large scale non-Gaussian linear dynamical system, tracking dominat ing directions of posterior covariance without imposing any factorization constr aints. Our approach can be scaled up to high-resolution images by reductions to numerical mathematics primitives and parallelization on several levels. In a fir st study, designs are found that improve significantly on others chosen independ ently for each slice or drawn at random.
****

Toward Provably Correct Feature Selection in Arbitrary Domains
Dimitris Margaritis
In this paper we address the problem of provably correct feature selection in ar
bitrary domains. An optimal solution to the problem is a Markov boundary, which
 is a minimal set of features that make the probability distribution of a target
 variable conditionally invariant to the state of all other features in the doma
in. While numerous algorithms for this problem have been proposed, their theore
tical correctness and practical behavior under arbitrary probability distributio
ns is unclear. We address this by introducing the Markov Boundary Theorem that
precisely characterizes the properties of an ideal Markov boundary, and use it t
o develop algorithms that learn a more general boundary that can capture complex
 interactions that only appear when the values of multiple features are consider
ed together. We introduce two algorithms: an exact, provably correct one as wel
l a more practical randomized anytime version, and show that they perform well o
n artificial as well as benchmark and real-world data sets. Throughout the pape
r we make minimal assumptions that consist of only a general set of axioms that
hold for every probability distribution, which gives these algorithms universal
applicability.
****
3D Object Recognition with Deep Belief Nets
Vinod Nair, Geoffrey E. Hinton
We introduce a new type of Deep Belief Net and evaluate it on a 3D object recogn
ition task. The top-level model is a third-order Boltzmann machine, trained usin
g a hybrid algorithm that combines both generative and discriminative gradients.
 Performance is evaluated on the NORB database(normalized-uniform version), whic
h contains stereo-pair images of objects under different lighting conditions and
 viewpoints. Our model achieves 6.5% error on the test set, which is close to th
e best published result for NORB (5.9%) using a convolutional neural net that ha
s built-in knowledge of translation invariance. It substantially outperforms sha
llow models such as SVMs (11.6%). DBNs are especially suited for semi-supervised
 learning, and to demonstrate this we consider a modified version of the NORB re
cognition task in which additional unlabeled images are created by applying smal
l translations to the images in the database. With the extra unlabeled data (and
 the same amount of labeled data as before), our model achieves 5.2% error, maki
ng it the current best result for NORB.
****
Improving Existing Fault Recovery Policies
Guy Shani, Christopher Meek
Automated recovery from failures is a key component in the management of large d
ata centers. Such systems typically employ a hand-made controller created by an
expert. While such controllers capture many important aspects of the recovery pr
ocess, they are often not systematically optimized to reduce costs such as serve
r downtime. In this paper we explain how to use data gathered from the interacti
ons of the hand-made controller with the system, to create an optimized controll
er. We suggest learning an indefinite horizon Partially Observable Markov Decisi
on Process, a model for decision making under uncertainty, and solve it using a
point-based algorithm. We describe the complete process, starting with data gath
ering, model learning, model checking procedures, and computing a policy. While
our paper focuses on a specific domain, our method is applicable to other system
s that use a hand-coded, imperfect controllers.
****
Convex Relaxation of Mixture Regression with Efficient Algorithms
Novi Quadrianto, John Lim, Dale Schuurmans, Tibério Caetano
We develop a convex relaxation of maximum a posteriori estimation of a mixture o
f regression models. Although our relaxation involves a semidefinite matrix vari
able, we reformulate the problem to eliminate the need for general semidefinite
programming. In particular, we provide two reformulations that admit fast algori
thms. The first is a max-min spectral reformulation exploiting quasi-Newton desc
ent. The second is a min-min reformulation consisting of fast alternating steps
of closed-form updates. We evaluate the methods against Expectation-Maximization

in a real problem of motion segmentation from video data.
****

Hierarchical Learning of Dimensional Biases in Human Categorization
Adam Sanborn, Nick Chater, Katherine A. Heller
Existing models of categorization typically represent to-be-classified items as points in a multidimensional space. While from a mathematical point of view, an infinite number of basis sets can be used to represent points in this space, the choice of basis set is psychologically crucial. People generally choose the same basis dimensions, and have a strong preference to generalize along the axes of these dimensions, but not diagonally". What makes some choices of dimension special? We explore the idea that the dimensions used by people echo the natural variation in the environment. Specifically, we present a rational model that does not assume dimensions, but learns the same type of dimensional generalizations that people display. This bias is shaped by exposing the model to many categories with a structure hypothesized to be like those which children encounter. Our model can be viewed as a type of transformed Dirichlet process mixture model, where it is the learning of the base distribution of the Dirichlet process which allows dimensional generalization.The learning behaviour of our model captures the developmental shift from roughly "isotropic" for children to the axis-aligned generalization that adults show."
****

Polynomial Semantic Indexing
Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Corinna Cortes, Mehryar Mohri
We present a class of nonlinear (polynomial) models that are discriminatively trained to directly map from the word content in a query-document or document-document pair to a ranking score. Dealing with polynomial models on word features is computationally challenging. We propose a low rank (but diagonal preserving) representation of our polynomial models to induce feasible memory and computation requirements. We provide an empirical study on retrieval tasks based on Wikipedia documents, where we obtain state-of-the-art performance while providing realistically scalable methods.
****

Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling
Lei Shi, Thomas Griffiths
The goal of perception is to infer the hidden states in the hierarchical process by which sensory data are generated. Human behavior is consistent with the optimal statistical solution to this problem in many tasks, including cue combination and orientation detection. Understanding the neural mechanisms underlying this behavior is of particular importance, since probabilistic computations are notoriously challenging. Here we propose a simple mechanism for Bayesian inference which involves averaging over a few feature detection neurons which fire at a rate determined by their similarity to a sensory stimulus. This mechanism is based on a Monte Carlo method known as importance sampling, commonly used in computer science and statistics. Moreover, a simple extension to recursive importance sampling can be used to perform hierarchical Bayesian inference. We identify a scheme for implementing importance sampling with spiking neurons, and show that this scheme can account for human behavior in cue combination and oblique effect.
****

Discriminative Network Models of Schizophrenia
Irina Rish, Benjamin Thyreau, Bertrand Thirion, Marion Plaze, Marie-laure Paillere-martinot, Catherine Martelli, Jean-luc Martinot, Jean-baptiste Poline, Guillermo Cecchi
Schizophrenia is a complex psychiatric disorder that has eluded a characterization in terms of local abnormalities of brain activity, and is hypothesized to affect the collective, ``emergent working of the brain. We propose a novel data-driven approach to capture emergent features using functional brain networks [Eguiluzet al] extracted from fMRI data, and demonstrate its advantage over traditional region-of-interest (ROI) and local, task-specific linear activation analyzes. Our results suggest that schizophrenia is indeed associated with disruption of g

lobal, emergent brain properties related to its functioning as a network, which cannot be explained by alteration of local activation patterns. Moreover, further exploitation of interactions by sparse Markov Random Field classifiers shows clear gain over linear methods, such as Gaussian Naive Bayes and SVM, allowing to reach 86% accuracy (over 50% baseline - random guess), which is quite remarkable given that it is based on a single fMRI experiment using a simple auditory task.

****

Sparsistent Learning of Varying-coefficient Models with Structural Changes

Mladen Kolar, Le Song, Eric Xing

To estimate the changing structure of a varying-coefficient varying-structure (VCVS) model remains an important and open problem in dynamic system modelling, which includes learning trajectories of stock prices, or uncovering the topology of an evolving gene network. In this paper, we investigate sparsistent learning of a sub-family of this model --- piecewise constant VCVS models. We analyze two main issues in this problem: inferring time points where structural changes occur and estimating model structure (i.e., model selection) on each of the constant segments. We propose a two-stage adaptive procedure, which first identifies jump points of structural changes and then identifies relevant covariates to a response on each of the segments. We provide an asymptotic analysis of the procedure, showing that with the increasing sample size, number of structural changes, and number of variables, the true model can be consistently selected. We demonstrate the performance of the method on synthetic data and apply it to the brain computer interface dataset. We also consider how this applies to structure estimation of time-varying probabilistic graphical models.

****

Dual Averaging Method for Regularized Stochastic Learning and Online Optimization

Lin Xiao

We consider regularized stochastic learning and online optimization problems, where the objective function is the sum of two convex terms: one is the loss function of the learning task, and the other is a simple regularization term such as L1-norm for sparsity. We develop a new online algorithm, the regularized dual averaging method, that can explicitly exploit the regularization structure in an online setting. In particular, at each iteration, the learning variables are adjusted by solving a simple optimization problem that involves the running average of all past subgradients of the loss functions and the whole regularization term, not just its subgradient. This method achieves the optimal convergence rate and often enjoys a low complexity per iteration similar as the standard stochastic gradient method. Computational experiments are presented for the special case of sparse online learning using L1-regularization.

****

Analysis of SVM with Indefinite Kernels

Yiming Ying, Colin Campbell, Mark Girolami

The recent introduction of indefinite SVM by Luss and dAspremont [15] has effectively demonstrated SVM classification with a non-positive semi-definite kernel (indefinite kernel). This paper studies the properties of the objective function introduced there. In particular, we show that the objective function is continuously differentiable and its gradient can be explicitly computed. Indeed, we further show that its gradient is Lipschitz continuous. The main idea behind our analysis is that the objective function is smoothed by the penalty term, in its saddle (min-max) representation, measuring the distance between the indefinite kernel matrix and the proxy positive semi-definite one. Our elementary result greatly facilitates the application of gradient-based algorithms. Based on our analysis, we further develop Nesterovs smooth optimization approach [16,17] for indefinite SVM which has an optimal convergence rate for smooth problems. Experiments on various benchmark datasets validate our analysis and demonstrate the efficiency of our proposed algorithms.

****

Quantification and the language of thought
Charles Kemp
Many researchers have suggested that the psychological complexity of a concept i s related to the length of its representation in a language of thought. As yet, however, there are few concrete proposals about the nature of this language. Th is paper makes one such proposal: the language of thought allows first order qua ntification (quantification over objects) more readily than second-order quantif ication (quantification over features). To support this proposal we present beha vioral results from a concept learning study inspired by the work of Shepard, Ho vland and Jenkins."
****

Exploring Functional Connectivities of the Human Brain using Multivariate Inform ation Analysis
Barry Chai, Dirk Walther, Diane Beck, Li Fei-fei
In this study, we present a method for estimating the mutual information for a l ocalized pattern of fMRI data. We show that taking a multivariate information ap proach to voxel selection leads to a decoding accuracy that surpasses an univari ate inforamtion approach and other standard voxel selection methods. Furthermore ,we extend the multivariate mutual information theory to measure the functional connectivity between distributed brain regions. By jointly estimating the inform ation shared by two sets of voxels we can reliably map out the connectivities in the human brain during experiment conditions. We validated our approach on a 6- way scene categorization fMRI experiment. The multivariate information analysis is able to ■nd strong information ■ow between PPA and RSC, which con■rms existin g neuroscience studies on scenes. Furthermore, by exploring over the whole brain , our method identifies other interesting ROIs that share information with the P PA, RSC scene network,suggesting interesting future work for neuroscientists.
****

Semi-supervised Learning using Sparse Eigenfunction Bases
Kaushik Sinha, Mikhail Belkin
We present a new framework for semi-supervised learning with sparse eigenfunctio n bases of kernel matrices. It turns out that when the \emph{cluster assumptio n} holds, that is, when the high density regions are sufficiently separated by low density valleys, each high density area corresponds to a unique representati ve eigenvector. Linear combination of such eigenvectors (or, more precisely, of their Nystrom extensions) provide good candidates for good classification functi ons. By first choosing an appropriate basis of these eigenvectors from unlabeled data and then using labeled data with Lasso to select a classifier in the span of these eigenvectors, we obtain a classifier, which has a very sparse represen tation in this basis. Importantly, the sparsity appears naturally from the clus ter assumption. Experimental results on a number of real-world data-sets show t hat our method is competitive with the state of the art semi-supervised learning algorithms and outperforms the natural base-line algorithm (Lasso in the Kernel PCA basis).
****

On Learning Rotations
Raman Arora
An algorithm is presented for online learning of rotations. The proposed algorit hm involves matrix exponentiated gradient updates and is motivated by the Von Ne umann divergence. The additive updates are skew-symmetric matrices with trace ze ro which comprise the Lie algebra of the rotation group. The orthogonality and u nit determinant of the matrix parameter are preserved using matrix logarithms a nd exponentials and the algorithm lends itself to interesting interpretations in terms of the computational topology of the compact Lie groups. The stability an d the computational complexity of the algorithm are discussed.
****

A Gaussian Tree Approximation for Integer Least-Squares
Jacob Goldberger, Amir Leshem
This paper proposes a new algorithm for the linear least squares problem where t he unknown variables are constrained to be in a finite set. The factor graph th

at corresponds to this problem is very loopy; in fact, it is a complete graph. H ence, applying the Belief Propagation (BP) algorithm yields very poor results. T he algorithm described here is based on an optimal tree approximation of the G aussian density of the unconstrained linear system. It is shown that even though the approximation is not directly applied to the exact discrete distribution, a pplying the BP algorithm to the modified factor graph outperforms current method s in terms of both performance and complexity. The improved performance of the p roposed algorithm is demonstrated on the problem of MIMO detection.
****

Nonlinear directed acyclic structure learning with weakly additive noise models
Arthur Gretton, Peter Spirtes, Robert Tillman
The recently proposed \emph{additive noise model} has advantages over previous s tructure learning algorithms, when attempting to recover some true data generati ng mechanism, since it (i) does not assume linearity or Gaussianity and (ii) can recover a unique DAG rather than an equivalence class. However, its original ex tension to the multivariate case required enumerating all possible DAGs, and for some special distributions, e.g. linear Gaussian, the model is invertible and t hus cannot be used for structure learning. We present a new approach which combi nes a PC style search using recent advances in kernel measures of conditional de pendence with local searches for additive noise models in substructures of the e quivalence class. This results in a more computationally efficient approach that is useful for arbitrary distributions even when additive noise models are inver tible. Experiments with synthetic and real data show that this method is more ac curate than previous methods when data are nonlinear and/or non-Gaussian.
****

FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs
Andrew McCallum, Karl Schultz, Sameer Singh
Discriminatively trained undirected graphical models have had wide empirical suc cess, and there has been increasing interest in toolkits that ease their applica tion to complex relational data. The power in relational models is in their rep eated structure and tied parameters; at issue is how to define these structures in a powerful and flexible way. Rather than using a declarative language, such a s SQL or first-order logic, we advocate using an imperative language to express various aspects of model structure, inference, and learning. By combining the t raditional, declarative, statistical semantics of factor graphs with imperative definitions of their construction and operation, we allow the user to mix declar ative and procedural domain knowledge, and also gain significant efficiencies. We have implemented such imperatively defined factor graphs in a system we call Factorie, a software library for an object-oriented, strongly-typed, functional language. In experimental comparisons to Markov Logic Networks on joint segment ation and coreference, we find our approach to be 3-15 times faster while reduci ng error by 20-25%-achieving a new state of the art.
****

Exponential Family Graph Matching and Ranking
James Petterson, Jin Yu, Julian Mcauley, Tibério Caetano
We present a method for learning max-weight matching predictors in bipartite gra phs. The method consists of performing maximum a posteriori estimation in expone ntial families with sufficient statistics that encode permutations and data feat ures. Although inference is in general hard, we show that for one very relevant application - document ranking - exact inference is efficient. For general model instances, an appropriate sampler is readily available. Contrary to existing ma x-margin matching models, our approach is statistically consistent and, in addit ion, experiments with increasing sample sizes indicate superior improvement over such models. We apply the method to graph matching in computer vision as well a s to a standard benchmark dataset for learning document ranking, in which we obt ain state-of-the-art results, in particular improving on max-margin variants. Th e drawback of this method with respect to max-margin alternatives is its runtime for large graphs, which is high comparatively.
****

Extending Phase Mechanism to Differential Motion Opponency for Motion Pop-out

Yicong Meng, Bertram Shi
We extend the concept of phase tuning, a ubiquitous mechanism in sensory neurons including motion and disparity detection neurons, to the motion contrast detection. We demonstrate that motion contrast can be detected by phase shifts between motion neuronal responses in different spatial regions. By constructing the differential motion opponency in response to motions in two different spatial regions, varying motion contrasts can be detected, where similar motion is detected by zero phase shifts and differences in motion by non-zero phase shifts. The model can exhibit either enhancement or suppression of responses by either different or similar motion in the surrounding. A primary advantage of the model is that the responses are selective to relative motion instead of absolute motion, which could model neurons found in neurophysiological experiments responsible for motion pop-out detection.
****

Periodic Step Size Adaptation for Single Pass On-line Learning
Chun-nan Hsu, Yu-ming Chang, Hanshen Huang, Yuh-jye Lee
It has been established that the second-order stochastic gradient descent (2SGD) method can potentially achieve generalization performance as well as empirical optimum in a single pass (i.e., epoch) through the training examples. However, 2SGD requires computing the inverse of the Hessian matrix of the loss function, which is prohibitively expensive. This paper presents Periodic Step-size Adaptation (PSA), which approximates the Jacobian matrix of the mapping function and explores a linear relation between the Jacobian and Hessian to approximate the Hessian periodically and achieve near-optimal results in experiments on a wide variety of models and tasks.
****

Construction of Nonparametric Bayesian Models from Parametric Bayes Equations
Peter Orbanz
We consider the general problem of constructing nonparametric Bayesian models on infinite-dimensional random objects, such as functions, infinite graphs or infinite permutations. The problem has generated much interest in machine learning, where it is treated heuristically, but has not been studied in full generality in nonparametric Bayesian statistics, which tends to focus on models over probability distributions. Our approach applies a standard tool of stochastic process theory, the construction of stochastic processes from their finite-dimensional marginal distributions. The main contribution of the paper is a generalization of the classic Kolmogorov extension theorem to conditional probabilities. This extension allows a rigorous construction of nonparametric Bayesian models from systems of finite-dimensional, parametric Bayes equations. Using this approach, we show (i) how existence of a conjugate posterior for the nonparametric model can be guaranteed by choosing conjugate finite-dimensional models in the construction, (ii) how the mapping to the posterior parameters of the nonparametric model can be explicitly determined, and (iii) that the construction of conjugate models in essence requires the finite-dimensional models to be in the exponential family. As an application of our constructive framework, we derive a model on infinite permutations, the nonparametric Bayesian analogue of a model recently proposed for the analysis of rank data.
****

Adaptive Regularization for Transductive Support Vector Machine
Zenglin Xu, Rong Jin, Jianke Zhu, Irwin King, Michael Lyu, Zhirong Yang
We discuss the framework of Transductive Support Vector Machine (TSVM) from the perspective of the regularization strength induced by the unlabeled data. In this framework, SVM and TSVM can be regarded as a learning machine without regularization and one with full regularization from the unlabeled data, respectively. Therefore, to supplement this framework of the regularization strength, it is necessary to introduce data-dependant partial regularization. To this end, we reformulate TSVM into a form with controllable regularization strength, which includes SVM and TSVM as special cases. Furthermore, we introduce a method of adaptive regularization that is data dependant and is based on the smoothness assumption.

Experiments on a set of benchmark data sets indicate the promising results of t
he proposed work compared with state-of-the-art TSVM algorithms.
****
Subject independent EEG-based BCI decoding
Siamac Fazli, Cristian Grozea, Marton Danoczy, Benjamin Blankertz, Florin Popesc
u, Klaus-Robert Müller
In the quest to make Brain Computer Interfacing (BCI) more usable, dry electrode
s have emerged that get rid of the initial 30 minutes required for placing an el
ectrode cap. Another time consuming step is the required individualized adaptat
ion to the BCI user, which involves another 30 minutes calibration for assessing
 a subjects brain signature. In this paper we aim to also remove this calibratio
n proceedure from BCI setup time by means of machine learning. In particular, we
 harvest a large database of EEG BCI motor imagination recordings (83 subjects)
for constructing a library of subject-specific spatio-temporal filters and deriv
e a subject independent BCI classifier. Our offline results indicate that BCI-na
{i}ve users could start real-time BCI use with no prior calibration at only a ve
ry moderate performance loss."
****
On the Convergence of the Concave-Convex Procedure
Gert Lanckriet, Bharath K. Sriperumbudur
The concave-convex procedure (CCCP) is a majorization-minimization algorithm tha
t solves d.c. (difference of convex functions) programs as a sequence of convex
programs. In machine learning, CCCP is extensively used in many learning algorit
hms like sparse support vector machines (SVMs), transductive SVMs, sparse princi
pal component analysis, etc. Though widely used in many applications, the conver
gence behavior of CCCP has not gotten a lot of specific attention. Yuille and Ra
ngarajan analyzed its convergence in their original paper, however, we believe t
he analysis is not complete. Although the convergence of CCCP can be derived fro
m the convergence of the d.c. algorithm (DCA), their proof is more specialized a
nd technical than actually required for the specific case of CCCP. In this paper
, we follow a different reasoning and show how Zangwills global convergence theo
ry of iterative algorithms provides a natural framework to prove the convergence
 of CCCP, allowing a more elegant and simple proof. This underlines Zangwills th
eory as a powerful and general framework to deal with the convergence issues of
iterative algorithms, after also being used to prove the convergence of algorith
ms like expectation-maximization, generalized alternating minimization, etc. In
this paper, we provide a rigorous analysis of the convergence of CCCP by address
ing these questions: (i) When does CCCP find a local minimum or a stationary poi
nt of the d.c. program under consideration? (ii) When does the sequence generate
d by CCCP converge? We also present an open problem on the issue of local conver
gence of CCCP.
****
Orthogonal Matching Pursuit From Noisy Random Measurements: A New Analysis
Sundeep Rangan, Alyson K. Fletcher
Orthogonal matching pursuit (OMP) is a widely used greedy algorithm for recoveri
ng sparse vectors from linear measurements. A well-known analysis of Tropp and
Gilbert shows that OMP can recover a k-sparse n-dimensional real vector from m =
 4k log(n) noise-free random linear measurements with a probability that goes to
 one as n goes to infinity. This work shows strengthens this result by showing t
hat a lower number of measurements, m = 2k log(n-k), is in fact sufficient for a
symptotic recovery. Moreover, this number of measurements is also sufficient for
 detection of the sparsity pattern (support) of the vector with measurement erro
rs provided the signal-to-noise ratio (SNR) scales to infinity. The scaling m =
2k log(n-k) exactly matches the number of measurements required by the more comp
lex lasso for signal recovery.
****
Noise Characterization, Modeling, and Reduction for In Vivo Neural Recording
Zhi Yang, Qi Zhao, Edward Keefer, Wentai Liu
Studying signal and noise properties of recorded neural data is critical in deve
loping more efficient algorithms to recover the encoded information. Important i

ssues exist in this research including the variant spectrum spans of neural spik es that make it difficult to choose a global optimal bandpass filter. Also, mult iple sources produce aggregated noise that deviates from the conventional white Gaussian noise. In this work, the spectrum variability of spikes is addressed, b ased on which the concept of adaptive bandpass filter that fits the spectrum of individual spikes is proposed. Multiple noise sources have been studied through analytical models as well as empirical measurements. The dominant noise source i s identified as neuron noise followed by interface noise of the electrode. This suggests that major efforts to reduce noise from electronics are not well spent. The measured noise from in vivo experiments shows a family of 1/f^{x} (x=1.5\pm 0.5) spectrum that can be reduced using noise shaping techniques. In summary, t he methods of adaptive bandpass filtering and noise shaping together result in s everal dB signal-to-noise ratio (SNR) enhancement.

****

Adaptive Regularization of Weight Vectors

Koby Crammer, Alex Kulesza, Mark Dredze

We present AROW, a new online learning algorithm that combines several propertie s of successful : large margin training, confidence weighting, and the capacity to handle non-separable data. AROW performs adaptive regularization of the predi ction function upon seeing each new instance, allowing it to perform especially well in the presence of label noise. We derive a mistake bound, similar in form to the second order perceptron bound, which does not assume separability. We al so relate our algorithm to recent confidence-weighted online learning techniques and empirically show that AROW achieves state-of-the-art performance and notabl e robustness in the case of non-separable data.

****

Posterior vs Parameter Sparsity in Latent Variable Models

Kuzman Ganchev, Ben Taskar, Fernando Pereira, João Gama

In this paper we explore the problem of biasing unsupervised models to favor spa rsity. We extend the posterior regularization framework [8] to encourage the mo del to achieve posterior sparsity on the unlabeled training data. We apply this new method to learn ■rst-order HMMs for unsupervised part-of-speech (POS) taggin g, and show that HMMs learned this way consistently and signi■cantly out-perform s both EM-trained HMMs, and HMMs with a sparsity-inducing Dirichlet prior traine d by variational EM. We evaluate these HMMs on three languages — English, Bulgar ian and Portuguese — under four conditions. We ■nd that our method always impro ves performance with respect to both baselines, while variational Bayes actually degrades performance in most cases. We increase accuracy with respect to EM by 2.5%-8.7% absolute and we see improvements even in a semisupervised condition wh ere a limited dictionary is provided.

****

Reconstruction of Sparse Circuits Using Multi-neuronal Excitation (RESCUME)

Tao Hu, Anthony Leonardo, Dmitri Chklovskii

One of the central problems in neuroscience is reconstructing synaptic connectiv ity in neural circuits. Synapses onto a neuron can be probed by sequentially sti mulating potentially pre-synaptic neurons while monitoring the membrane voltage of the post-synaptic neuron. Reconstructing a large neural circuit using such a “brute force” approach is rather time-consuming and inefficient because the conn ectivity in neural circuits is sparse. Instead, we propose to measure a post-syn aptic neuron’s voltage while stimulating simultaneously multiple randomly chosen potentially pre-synaptic neurons. To extract the weights of individual synaptic connections we apply a decoding algorithm recently developed for compressive se nsing. Compared to the brute force approach, our method promises significant tim e savings that grow with the size of the circuit. We use computer simulations to find optimal stimulation parameters and explore the feasibility of our reconstr uction method under realistic experimental conditions including noise and non-li near synaptic integration. Multiple-neuron stimulation allows reconstructing syn aptic connectivity just from the spiking activity of post-synaptic neurons, even when sub-threshold voltage is unavailable. By using calcium indicators, voltage -sensitive dyes, or multi-electrode arrays one could monitor activity of multipl

e post-synaptic neurons simultaneously, thus mapping their synaptic inputs in pa
rallel, potentially reconstructing a complete neural circuit.
****

Label Selection on Graphs
Andrew Guillory, Jeff A. Bilmes
We investigate methods for selecting sets of labeled vertices for use in predict
ing the labels of vertices on a graph. We specifically study methods which choo
se a single batch of labeled vertices (i.e. offline, non sequential methods). I
n this setting, we find common graph smoothness assumptions directly motivate si
mple label selection methods with interesting theoretical guarantees. These meth
ods bound prediction error in terms of the smoothness of the true labels with re
spect to the graph. Some of these bounds give new motivations for previously pr
oposed algorithms, and some suggest new algorithms which we evaluate. We show i
mproved performance over baseline methods on several real world data sets.
****

A Fast, Consistent Kernel Two-Sample Test
Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, Bharath K. Sriperumbudur
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Robust Nonparametric Regression with Metric-Space Valued Output
Matthias Hein
Motivated by recent developments in manifold-valued regression we propose a fami
ly of nonparametric kernel-smoothing estimators with metric-space valued output
including a robust median type estimator and the classical Frechet mean. Dependi
ng on the choice of the output space and the chosen metric the estimator reduces
 to partially well-known procedures for multi-class classification, multivariate
 regression in Euclidean space, regression with manifold-valued output and even
some cases of structured output learning. In this paper we focus on the case of
regression with manifold-valued input and output. We show pointwise and Bayes co
nsistency for all estimators in the family for the case of manifold-valued outpu
t and illustrate the robustness properties of the estimator with experiments.
****

Kernels and learning curves for Gaussian process regression on random graphs
Peter Sollich, Matthew Urry, Camille Coti
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Thresholding Procedures for High Dimensional Variable Selection and Statistical
Estimation
Shuheng Zhou
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Beyond Categories: The Visual Memex Model for Reasoning About Object Relationshi
ps
Tomasz Malisiewicz, Alyosha Efros
The use of context is critical for scene understanding in computer vision, where
 the recognition of an object is driven by both local appearance and the objects
 relationship to other elements of the scene (context). Most current approaches
 rely on modeling the relationships between object categories as a source of con
text. In this paper we seek to move beyond categories to provide a richer appear
ance-based model of context. We present an exemplar-based model of objects and
their relationships, the Visual Memex, that encodes both local appearance and 2D

spatial context between object instances. We evaluate our model on Torralbas pr
oposed Context Challenge against a baseline category-based system. Our experimen
ts suggest that moving beyond categories for context modeling appears to be quit
e beneficial, and may be the critical missing ingredient in scene understanding
systems.
****

Conditional Random Fields with High-Order Features for Sequence Labeling
Nan Ye, Wee Lee, Hai Chieu, Dan Wu
Dependencies among neighbouring labels in a sequence is an important source of i
nformation for sequence labeling problems. However, only dependencies between ad
jacent labels are commonly exploited in practice because of the high computation
al complexity of typical inference algorithms when longer distance dependencies
are taken into account. In this paper, we show that it is possible to design eff
icient inference algorithms for a conditional random field using features that d
epend on long consecutive label sequences (high-order features), as long as the
number of distinct label sequences in the features used is small. This leads to
efficient learning algorithms for these conditional random fields. We show exper
imentally that exploiting dependencies using high-order features can lead to sub
stantial performance improvements for some problems and discuss conditions under
which high-order features can be effective.
****

Abstraction and Relational learning
Charles Kemp, Alan Jern
Many categories are better described by providing relational information than li
sting characteristic features. We present a hierarchical generative model that
helps to explain how relational categories are learned and used. Our model learn
s abstract schemata that specify the relational similarities shared by members o
f a category, and our emphasis on abstraction departs from previous theoretical
proposals that focus instead on comparison of concrete instances. Our first expe
riment suggests that our abstraction-based account can address some of the tasks
that have previously been used to support comparison-based approaches. Our seco
nd experiment focuses on one-shot schema learning, a problem that raises challen
ges for comparison-based approaches but is handled naturally by our abstraction-
based account.
****

Fast Graph Laplacian Regularized Kernel Learning via Semidefinite–Quadratic–Line
ar Programming
Xiao-ming Wu, Anthony So, Zhenguo Li, Shuo-yen Li
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

A Data-Driven Approach to Modeling Choice
Vivek Farias, Srikanth Jagabathula, Devavrat Shah
We visit the following fundamental problem: For a `generic model of consumer cho
ice (namely, distributions over preference lists) and a limited amount of data o
n how consumers actually make decisions (such as marginal preference information
), how may one predict revenues from offering a particular assortment of choices
? This problem is central to areas within operations research, marketing and eco
nometrics. We present a framework to answer such questions and design a number o
f tractable algorithms (from a data and computational standpoint) for the same.
****

Anomaly Detection with Score functions based on Nearest Neighbor Graphs
Manqi Zhao, Venkatesh Saligrama
We propose a novel non-parametric adaptive anomaly detection algorithm for high
dimensional data based on score functions derived from nearest neighbor graphs o
n n-point nominal data. Anomalies are declared whenever the score of a test samp
le falls below q, which is supposed to be the desired false alarm level. The res
ulting anomaly detector is shown to be asymptotically optimal in that it is unif

ormly most powerful for the specified false alarm level, q, for the case when th
e anomaly density is a mixture of the nominal and a known density. Our algorithm
 is computationally efficient, being linear in dimension and quadratic in data s
ize. It does not require choosing complicated tuning parameters or function appr
oximation classes and it can adapt to local structure such as local change in di
mensionality. We demonstrate the algorithm on both artificial and real data sets
 in high dimensional feature spaces.
****

Submodularity Cuts and Applications
Yoshinobu Kawahara, Kiyohito Nagano, Koji Tsuda, Jeff A. Bilmes
Several key problems in machine learning, such as feature selection and active l
earning, can be formulated as submodular set function maximization. We present
herein a novel algorithm for maximizing a submodular set function under a cardin
ality constraint --- the algorithm is based on a cutting-plane method and is imp
lemented as an iterative small-scale binary-integer linear programming procedure
. It is well known that this problem is NP-hard, and the approximation factor ac
hieved by the greedy algorithm is the theoretical limit for polynomial time. As
for (non-polynomial time) exact algorithms that perform reasonably in practice,
there has been very little in the literature although the problem is quite impor
tant for many applications. Our algorithm is guaranteed to find the exact soluti
on in finite iterations, and it converges fast in practice due to the efficiency
 of the cutting-plane mechanism. Moreover, we also provide a method that produce
s successively decreasing upper-bounds of the optimal solution, while our algori
thm provides successively increasing lower-bounds. Thus, the accuracy of the cu
rrent solution can be estimated at any point, and the algorithm can be stopped e
arly once a desired degree of tolerance is met. We evaluate our algorithm on se
nsor placement and feature selection applications showing good performance.
****

Bayesian Sparse Factor Models and DAGs Inference and Comparison
Ricardo Henao, Ole Winther
In this paper we present a novel approach to learn directed acyclic graphs (DAG)
 and factor models within the same framework while also allowing for model compa
rison between them. For this purpose, we exploit the connection between factor m
odels and DAGs to propose Bayesian hierarchies based on spike and slab priors to
 promote sparsity, heavy-tailed priors to ensure identifiability and predictive
densities to perform the model comparison. We require identifiability to be able
 to produce variable orderings leading to valid DAGs and sparsity to learn the s
tructures. The effectiveness of our approach is demonstrated through extensive e
xperiments on artificial and biological data showing that our approach outperfor
m a number of state of the art methods.
****

A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds
Paris Smaragdis, Madhusudana Shashanka, Bhiksha Raj
In this paper we present an algorithm for separating mixed sounds from a monoph
onic recording. Our approach makes use of training data which allows us to lear
n representations of the types of sounds that compose the mixture. In contrast
to popular methods that attempt to extract com- pact generalizable models for e
ach sound from training data, we employ the training data itself as a represent
ation of the sources in the mixture. We show that mixtures of known sounds can
be described as sparse com- binations of the training data itself, and in doing
 so produce signi■cantly better separation results as compared to similar syste
ms based on compact statistical models.
****

Data-driven calibration of linear estimators with minimal penalties
Sylvain Arlot, Francis Bach
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

fMRI-Based Inter-Subject Cortical Alignment Using Functional Connectivity
Bryan Conroy, Ben Singer, James Haxby, Peter J. Ramadge
The inter-subject alignment of functional MRI (fMRI) data is important for improving the statistical power of fMRI group analyses. In contrast to existing anatomically-based methods, we propose a novel multi-subject algorithm that derives a functional correspondence by aligning spatial patterns of functional connectivity across a set of subjects. We test our method on fMRI data collected during a movie viewing experiment. By cross-validating the results of our algorithm, we show that the correspondence successfully generalizes to a secondary movie dataset not used to derive the alignment.
****
Unsupervised feature learning for audio classification using convolutional deep belief networks
Honglak Lee, Peter Pham, Yan Largman, Andrew Ng
In recent years, deep learning approaches have gained significant interest as a way of building hierarchical representations from unlabeled data. However, to our knowledge, these deep learning approaches have not been extensively studied for auditory data. In this paper, we apply convolutional deep belief networks to audio data and empirically evaluate them on various audio classification tasks. For the case of speech data, we show that the learned features correspond to phones/phonemes. In addition, our feature representations trained from unlabeled audio data show very good performance for multiple audio classification tasks. We hope that this paper will inspire more research on deep learning approaches applied to a wide range of audio recognition tasks.
****
Accelerating Bayesian Structural Inference for Non-Decomposable Gaussian Graphical Models
Baback Moghaddam, Emtiyaz Khan, Kevin P. Murphy, Benjamin M. Marlin
In this paper we make several contributions towards accelerating approximate Bayesian structural inference for non-decomposable GGMs. Our first contribution is to show how to efficiently compute a BIC or Laplace approximation to the marginal likelihood of non-decomposable graphs using convex methods for precision matrix estimation. This optimization technique can be used as a fast scoring function inside standard Stochastic Local Search (SLS) for generating posterior samples. Our second contribution is a novel framework for efficiently generating large sets of high-quality graph topologies without performing local search. This graph proposal method, which we call Neighborhood Fusion" (NF), samples candidate Markov blankets at each node using sparse regression techniques. Our final contribution is a hybrid method combining the complementary strengths of NF and SLS. Experimental results in structural recovery and prediction tasks demonstrate that NF and hybrid NF/SLS out-perform state-of-the-art local search methods, on both synthetic and real-world datasets, when realistic computational limits are imposed."
****
Learning transport operators for image manifolds
Benjamin Culpepper, Bruno Olshausen
We describe a method for learning a group of continuous transformation operators to traverse smooth nonlinear manifolds. The method is applied to model how natural images change over time and scale. The group of continuous transform operators is represented by a basis that is adapted to the statistics of the data so that the in■nitesimal generator for a measurement orbit can be produced by a linear combination of a few basis elements. We illustrate how the method can be used to ef■ciently code time-varying images by describing changes across time and scale in terms of the learned operators.
****
Manifold Embeddings for Model-Based Reinforcement Learning under Partial Observability
Keith Bush, Joelle Pineau
Interesting real-world datasets often exhibit nonlinear, noisy, continuous-valued states that are unexplorable, are poorly described by first principles, and ar

e only partially observable. If partial observability can be overcome, these con straints suggest the use of model-based reinforcement learning. We experiment w ith manifold embeddings as the reconstructed observable state-space of an off-li ne, model-based reinforcement learning approach to control. We demonstrate the e mbedding of a system changes as a result of learning and that the best performin g embeddings well-represent the dynamics of both the uncontrolled and adaptively controlled system. We apply this approach in simulation to learn a neurostimul ation policy that is more efficient in treating epilepsy than conventional polic ies. We then demonstrate the learned policy completely suppressing seizures in real-world neurostimulation experiments on actual animal brain slices.
****

Ensemble Nystrom Method
Sanjiv Kumar, Mehryar Mohri, Ameet Talwalkar
A crucial technique for scaling kernel methods to very large data sets reaching or exceeding millions of instances is based on low-rank approximation of kernel matrices. We introduce a new family of algorithms based on mixtures of Nystrom a pproximations, ensemble Nystrom algorithms, that yield more accurate low-rank ap proximations than the standard Nystrom method. We give a detailed study of multi ple variants of these algorithms based on simple averaging, an exponential weigh t method, or regression-based methods. We also present a theoretical analysis of these algorithms, including novel error bounds guaranteeing a better convergenc e rate than the standard Nystrom method. Finally, we report the results of exten sive experiments with several data sets containing up to 1M points demonstrating the signi■cant performance improvements gained over the standard Nystrom approx imation.
****

Manifold Regularization for SIR with Rate Root-n Convergence
Wei Bian, Dacheng Tao
In this paper, we study the manifold regularization for the Sliced Inverse Regre ssion (SIR). The manifold regularization improves the standard SIR in two aspect s: 1) it encodes the local geometry for SIR and 2) it enables SIR to deal with t ransductive and semi-supervised learning problems. We prove that the proposed gr aph Laplacian based regularization is convergent at rate root-n. The projection directions of the regularized SIR are optimized by using a conjugate gradient me thod on the Grassmann manifold. Experimental results support our theory.
****

STDP enables spiking neurons to detect hidden causes of their inputs
Bernhard Nessler, Michael Pfeiffer, Wolfgang Maass
The principles by which spiking neurons contribute to the astounding computation al power of generic cortical microcircuits, and how spike-timing-dependent plast icity (STDP) of synaptic weights could generate and maintain this computational function, are unknown. We show here that STDP, in conjunction with a stochastic soft winner-take-all (WTA) circuit, induces spiking neurons to generate through their synaptic weights implicit internal models for subclasses (or causes") of the high-dimensional spike patterns of hundreds of pre-synaptic neurons. Hence t hese neurons will fire after learning whenever the current input best matches t heir internal model. The resulting computational function of soft WTA circuits, a common network motif of cortical microcircuits, could therefore be a drastic d imensionality reduction of information streams, together with the autonomous cre ation of internal models for the probability distributions of their input patter ns. We show that the autonomous generation and maintenance of this computational function can be explained on the basis of rigorous mathematical principles. In particular, we show that STDP is able to approximate a stochastic online Expecta tion-Maximization (EM) algorithm for modeling the input data. A corresponding re sult is shown for Hebbian learning in artificial neural networks."
****

Locality-sensitive binary codes from shift-invariant kernels
Maxim Raginsky, Svetlana Lazebnik
This paper addresses the problem of designing binary codes for high-dimensional data such that vectors that are similar in the original space map to similar bin

ary strings. We introduce a simple distribution-free encoding scheme based on ra
ndom projections, such that the expected Hamming distance between the binary cod
es of two vectors is related to the value of a shift-invariant kernel (e.g., a G
aussian kernel) between the vectors. We present a full theoretical analysis of t
he convergence properties of the proposed scheme, and report favorable experimen
tal performance as compared to a recent state-of-the-art method, spectral hashin
g.
****

Regularized Distance Metric Learning:Theory and Algorithm
Rong Jin, Shijun Wang, Yang Zhou
In this paper, we examine the generalization error of regularized distance metri
c learning. We show that with appropriate constraints, the generalization error
of regularized distance metric learning could be independent from the dimensiona
lity, making it suitable for handling high dimensional data. In addition, we pre
sent an efficient online learning algorithm for regularized distance metric lear
ning. Our empirical studies with data classification and face recognition show t
hat the proposed algorithm is (i) effective for distance metric learning when co
mpared to the state-of-the-art methods, and (ii) efficient and robust for high d
imensional data.
****

Time-Varying Dynamic Bayesian Networks
Le Song, Mladen Kolar, Eric Xing
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

A General Projection Property for Distribution Families
Yao-liang Yu, Yuxi Li, Dale Schuurmans, Csaba Szepesvári
We prove that linear projections between distribution families with fixed first
and second moments are surjective, regardless of dimension. We further extend th
is result to families that respect additional constraints, such as symmetry, uni
modality and log-concavity. By combining our results with classic univariate ine
qualities, we provide new worst-case analyses for natural risk criteria arising
in different fields. One discovery is that portfolio selection under the worst-c
ase value-at-risk and conditional value-at-risk criteria yields identical portfo
lios.
****

Region-based Segmentation and Object Detection
Stephen Gould, Tianshi Gao, Daphne Koller
Object detection and multi-class image segmentation are two closely related task
s that can be greatly improved when solved jointly by feeding information from o
ne task to the other. However, current state-of-the-art models use a separate re
presentation for each task making joint inference clumsy and leaving classificat
ion of many parts of the scene ambiguous. In this work, we propose a hierarchica
l region-based approach to joint object detection and image segmentation. Our ap
proach reasons about pixels, regions and objects in a coherent probabilistic mod
el. Importantly, our model gives a single unified description of the scene. We e
xplain every pixel in the image and enforce global consistency between all varia
bles in our model. We run experiments on challenging vision datasets and show si
gnificant improvement over state-of-the-art object detection accuracy.
****

Hierarchical Mixture of Classification Experts Uncovers Interactions between Bra
in Regions
Bangpeng Yao, Dirk Walther, Diane Beck, Li Fei-fei
The human brain can be described as containing a number of functional regions. F
or a given task, these regions, as well as the connections between them, play a
key role in information processing in the brain. However, most existing multi-vo
xel pattern analysis approaches either treat multiple functional regions as one
large uniform region or several independent regions, ignoring the connections be

tween regions. In this paper, we propose to model such connections in an Hidden Conditional Random Field (HCRF) framework, where the classifier of one region of interest (ROI) makes predictions based on not only its voxels but also the classifier predictions from ROIs that it connects to. Furthermore, we propose a structural learning method in the HCRF framework to automatically uncover the connections between ROIs. Experiments on fMRI data acquired while human subjects viewing images of natural scenes show that our model can improve the top-level (the classifier combining information from all ROIs) and ROI-level prediction accuracy, as well as uncover some meaningful connections between ROIs.
****

Unsupervised Detection of Regions of Interest Using Iterative Link Analysis
Gunhee Kim, Antonio Torralba
This paper proposes a fast and scalable alternating optimization technique to detect regions of interest (ROIs) in cluttered Web images without labels. The proposed approach discovers highly probable regions of object instances by iteratively repeating the following two functions: (1) choose the exemplar set (i.e. small number of high ranked reference ROIs) across the dataset and (2) refine the ROIs of each image with respect to the exemplar set. These two subproblems are formulated as ranking in two different similarity networks of ROI hypotheses by link analysis. The experiments with the PASCAL 06 dataset show that our unsupervised localization performance is better than one of state-of-the-art techniques and comparable to supervised methods. Also, we test the scalability of our approach with five objects in Flickr dataset consisting of more than 200,000 images.
****

Fast Learning from Non-i.i.d. Observations
Ingo Steinwart, Andreas Christmann
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.
****

Evaluating multi-class learning strategies in a generative hierarchical framework for object detection
Sanja Fidler, Marko Boben, Ales Leonardis
Multiple object class learning and detection is a challenging problem due to the large number of object classes and their high visual variability. Specialized detectors usually excel in performance, while joint representations optimize sharing and reduce inference time --- but are complex to train. Conveniently, sequential learning of categories cuts down training time by transferring existing knowledge to novel classes, but cannot fully exploit the richness of shareability and might depend on ordering in learning. In hierarchical frameworks these issues have been little explored. In this paper, we show how different types of multi-class learning can be done within one generative hierarchical framework and provide a rigorous experimental analysis of various object class learning strategies as the number of classes grows. Specifically, we propose, evaluate and compare three important types of multi-class learning: 1.) independent training of individual categories, 2.) joint training of classes, 3.) sequential learning of classes. We explore and compare their computational behavior (space and time) and detection performance as a function of the number of learned classes on several recognition data sets.
****

Optimal Scoring for Unsupervised Learning
Zhihua Zhang, Guang Dai
We are often interested in casting classification and clustering problems in a regression framework, because it is feasible to achieve some statistical properties in this framework by imposing some penalty criteria. In this paper we illustrate optimal scoring, which was originally proposed for performing Fisher linear discriminant analysis by regression, in the application of unsupervised learning. In particular, we devise a novel clustering algorithm that we call optimal discriminant clustering (ODC). We associate our algorithm with the existing unsuper

vised learning algorithms such as spectral clustering, discriminative clustering and sparse principal component analysis. Thus, our work shows that optimal scoring provides a new approach to the implementation of unsupervised learning. This approach facilitates the development of new unsupervised learning algorithms.
****

Matrix Completion from Noisy Entries
Raghunandan Keshavan, Andrea Montanari, Sewoong Oh
Given a matrix M of low-rank, we consider the problem of reconstructing it from noisy observations of a small, random subset of its entries. The problem arises in a variety of applications, from collaborative filtering (the 'Netflix problem') to structure-from-motion and positioning. We study a low complexity algorithm introduced in [1], based on a combination of spectral techniques and manifold optimization, that we call here OPTSPACE. We prove performance guarantees that are order-optimal in a number of circumstances.
****

A Generalized Natural Actor-Critic Algorithm
Tetsuro Morimura, Eiji Uchibe, Junichiro Yoshimoto, Kenji Doya
Policy gradient Reinforcement Learning (RL) algorithms have received much attention in seeking stochastic policies that maximize the average rewards. In addition, extensions based on the concept of the Natural Gradient (NG) show promising learning efficiency because these regard metrics for the task. Though there are two candidate metrics, Kakades Fisher Information Matrix (FIM) and Morimuras FIM, all RL algorithms with NG have followed the Kakades approach. In this paper, we describe a generalized Natural Gradient (gNG) by linearly interpolating the two FIMs and propose an efficient implementation for the gNG learning based on a theory of the estimating function, generalized Natural Actor-Critic (gNAC). The gNAC algorithm involves a near optimal auxiliary function to reduce the variance of the gNG estimates. Interestingly, the gNAC can be regarded as a natural extension of the current state-of-the-art NAC algorithm, as long as the interpolating parameter is appropriately selected. Numerical experiments showed that the proposed gNAC algorithm can estimate gNG efficiently and outperformed the NAC algorithm.
****

Efficient Moments-based Permutation Tests
Chunxiao Zhou, Huixia Wang, Yongmei Wang
In this paper, we develop an efficient moments-based permutation test approach to improve the system's efficiency by approximating the permutation distribution of the test statistic with Pearson distribution series. This approach involves the calculation of the first four moments of the permutation distribution. We propose a novel recursive method to derive these moments theoretically and analytically without any permutation. Experimental results using different test statistics are demonstrated using simulated data and real data. The proposed strategy takes advantage of nonparametric permutation tests and parametric Pearson distribution approximation to achieve both accuracy and efficiency.
****

Maximum likelihood trajectories for continuous-time Markov chains
Theodore Perkins
Continuous-time Markov chains are used to model systems in which transitions between states as well as the time the system spends in each state are random. Many computational problems related to such chains have been solved, including determining state distributions as a function of time, parameter estimation, and control. However, the problem of inferring most likely trajectories, where a trajectory is a sequence of states as well as the amount of time spent in each state, appears unsolved. We study three versions of this problem: (i) an initial value problem, in which an initial state is given and we seek the most likely trajectory until a given final time, (ii) a boundary value problem, in which initial and final states and times are given, and we seek the most likely trajectory connecting them, and (iii) trajectory inference under partial observability, analogous to finding maximum likelihood trajectories for hidden Markov models. We show that maximum likelihood trajectories are not always well-defined, and describe

a polynomial time test for well-definedness. When well-definedness holds, we sh
ow that each of the three problems can be solved in polynomial time, and we deve
lop efficient dynamic programming algorithms for doing so.
****

Lower bounds on minimax rates for nonparametric regression with additive sparsit
y and smoothness
Garvesh Raskutti, Bin Yu, Martin J. Wainwright
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Sufficient Conditions for Agnostic Active Learnable
Liwei Wang
We study pool-based active learning in the presence of noise, i.e. the agnostic
setting. Previous works have shown that the effectiveness of agnostic active lea
rning depends on the learning problem and the hypothesis space. Although there a
re many cases on which active learning is very useful, it is also easy to constr
uct examples that no active learning algorithm can have advantage. In this paper
, we propose intuitively reasonable sufficient conditions under which agnostic a
ctive learning algorithm is strictly superior to passive supervised learning. We
 show that under some noise condition, if the classification boundary and the un
derlying distribution are smooth to a finite order, active learning achieves pol
ynomial improvement in the label complexity; if the boundary and the distributio
n are infinitely smooth, the improvement is exponential.
****

The Ordered Residual Kernel for Robust Motion Subspace Clustering
Tat-jun Chin, Hanzi Wang, David Suter
We present a novel and highly effective approach for multi-body motion segmentat
ion. Drawing inspiration from robust statistical model fitting, we estimate puta
tive subspace hypotheses from the data. However, instead of ranking them we enca
psulate the hypotheses in a novel Mercer kernel which elicits the potential of t
wo point trajectories to have emerged from the same subspace. The kernel permits
 the application of well-established statistical learning methods for effective
outlier rejection, automatic recovery of the number of motions and accurate segm
entation of the point trajectories. The method operates well under severe outlie
rs arising from spurious trajectories or mistracks. Detailed experiments on a re
cent benchmark dataset (Hopkins 155) show that our method is superior to other s
tate-of-the-art approaches in terms of recovering the number of motions, segment
ation accuracy, robustness against gross outliers and computational efficiency.
****

Learning to Rank by Optimizing NDCG Measure
Hamed Valizadegan, Rong Jin, Ruofei Zhang, Jianchang Mao
Learning to rank is a relatively new field of study, aiming to learn a ranking f
unction from a set of training data with relevancy labels. The ranking algorithm
s are often evaluated using Information Retrieval measures, such as Normalized D
iscounted Cumulative Gain [1] and Mean Average Precision [2]. Until recently, mo
st learning to rank algorithms were not using a loss function related to the abo
ve mentioned evaluation measures. The main difficulty in direct optimization of
these measures is that they depend on the ranks of documents, not the numerical
values output by the ranking function. We propose a probabilistic framework that
 addresses this challenge by optimizing the expectation of NDCG over all the pos
sible permutations of documents. A relaxation strategy is used to approximate th
e average of NDCG over the space of permutation, and a bound optimization approa
ch is proposed to make the computation efficient. Extensive experiments show tha
t the proposed algorithm outperforms state-of-the-art ranking algorithms on seve
ral benchmark data sets.
****

Slow Learners are Fast
Martin Zinkevich, John Langford, Alex Smola

Online learning algorithms have impressive convergence properties when it comes to risk minimization and convex games on very large problems. However, they are inherently sequential in their design which prevents them from taking advantage of modern multi-core architectures. In this paper we prove that online learning with delayed updates converges well, thereby facilitating parallel online learning.

****

Tracking Dynamic Sources of Malicious Activity at Internet Scale
Shobha Venkataraman, Avrim Blum, Dawn Song, Subhabrata Sen, Oliver Spatscheck

We formulate and address the problem of discovering dynamic malicious regions on the Internet. We model this problem as one of adaptively pruning a known decision tree, but with additional challenges: (1) severe space requirements, since the underlying decision tree has over 4 billion leaves, and (2) a changing target function, since malicious activity on the Internet is dynamic. We present a novel algorithm that addresses this problem, by putting together a number of different ``experts algorithms and online paging algorithms. We prove guarantees on our algorithms performance as a function of the best possible pruning of a similar size, and our experiments show that our algorithm achieves high accuracy on large real-world data sets, with significant improvements over existing approaches.

****

Graph Zeta Function in the Bethe Free Energy and Loopy Belief Propagation
Yusuke Watanabe, Kenji Fukumizu

We propose a new approach to the analysis of Loopy Belief Propagation (LBP) by establishing a formula that connects the Hessian of the Bethe free energy with the edge zeta function. The formula has a number of theoretical implications on LBP. It is applied to give a sufficient condition that the Hessian of the Bethe free energy is positive definite, which shows non-convexity for graphs with multiple cycles. The formula clarifies the relation between the local stability of a fixed point of LBP and local minima of the Bethe free energy. We also propose a new approach to the uniqueness of LBP fixed point, and show various conditions of uniqueness.

****

Learning in Markov Random Fields using Tempered Transitions
Russ R. Salakhutdinov

Markov random fields (MRFs), or undirected graphical models, provide a powerful framework for modeling complex dependencies among random variables. Maximum likelihood learning in MRFs is hard due to the presence of the global normalizing constant. In this paper we consider a class of stochastic approximation algorithms of Robbins-Monro type that uses Markov chain Monte Carlo to do approximate maximum likelihood learning. We show that using MCMC operators based on tempered transitions enables the stochastic approximation algorithm to better explore highly multimodal distributions, which considerably improves parameter estimates in large densely-connected MRFs. Our results on MNIST and NORB datasets demonstrate that we can successfully learn good generative models of high-dimensional, richly structured data and perform well on digit and object recognition tasks.

****

Statistical Consistency of Top-k Ranking
Fen Xia, Tie-yan Liu, Hang Li

This paper is concerned with the consistency analysis on listwise ranking methods. Among various ranking methods, the listwise methods have competitive performances on benchmark datasets and are regarded as one of the state-of-the-art approaches. Most listwise ranking methods manage to optimize ranking on the whole list (permutation) of objects, however, in practical applications such as information retrieval, correct ranking at the top k positions is much more important. This paper aims to analyze whether existing listwise ranking methods are statistically consistent in the top-k setting. For this purpose, we define a top-k ranking framework, where the true loss (and thus the risks) are defined on the basis of top-k subgroup of permutations. This framework can include the permutation-level ranking framework proposed in previous work as a special case. Based on the new framework, we derive sufficient conditions for a listwise ranking method to be

consistent with the top-k true loss, and show an effective way of modifying the surrogate loss functions in existing methods to satisfy these conditions. Experimental results show that after the modifications, the methods can work significantly better than their original versions.
****

Speaker Comparison with Inner Product Discriminant Functions
Zahi Karam, Douglas Sturim, William Campbell
Speaker comparison, the process of finding the speaker similarity between two speech signals, occupies a central role in a variety of applications---speaker verification, clustering, and identification. Speaker comparison can be placed in a geometric framework by casting the problem as a model comparison process. For a given speech signal, feature vectors are produced and used to adapt a Gaussian mixture model (GMM). Speaker comparison can then be viewed as the process of compensating and finding metrics on the space of adapted models. We propose a framework, inner product discriminant functions (IPDFs), which extends many common techniques for speaker comparison: support vector machines, joint factor analysis, and linear scoring. The framework uses inner products between the parameter vectors of GMM models motivated by several statistical methods. Compensation of nuisances is performed via linear transforms on GMM parameter vectors. Using the IPDF framework, we show that many current techniques are simple variations of each other. We demonstrate, on a 2006 NIST speaker recognition evaluation task, new scoring methods using IPDFs which produce excellent error rates and require significantly less computation than current techniques.
****

Asymptotic Analysis of MAP Estimation via the Replica Method and Compressed Sensing
Sundeep Rangan, Vivek Goyal, Alyson K. Fletcher
The replica method is a non-rigorous but widely-used technique from statistical physics used in the asymptotic analysis of many large random nonlinear problems. This paper applies the replica method to non-Gaussian MAP estimation. It is shown that with large random linear measurements and Gaussian noise, the asymptotic behavior of the MAP estimate of an n-dimensional vector ``decouples as n scalar MAP estimators. The result is a counterpart to Guo and Verdus replica analysis on MMSE estimation. The replica MAP analysis can be readily applied to many estimators used in compressed sensing, including basis pursuit, lasso, linear estimation with thresholding and zero-norm estimation. In the case of lasso estimation, the scalar estimator reduces to a soft-thresholding operator and for zero-norm estimation it reduces to a hard-threshold. Among other benefits, the replica method provides a computationally tractable method for exactly computing various performance metrics including MSE and sparsity recovery.
****

Statistical Models of Linear and Nonlinear Contextual Interactions in Early Visual Processing
Ruben Coen-cagli, Peter Dayan, Odelia Schwartz
A central hypothesis about early visual processing is that it represents inputs in a coordinate system matched to the statistics of natural scenes. Simple versions of this lead to Gabor-like receptive fields and divisive gain modulation from local surrounds; these have led to influential neural and psychological models of visual processing. However, these accounts are based on an incomplete view of the visual context surrounding each point. Here, we consider an approximate model of linear and non-linear correlations between the responses of spatially distributed Gabor-like receptive fields, which, when trained on an ensemble of natural scenes, unifies a range of spatial context effects. The full model accounts for neural surround data in primary visual cortex (V1), provides a statistical foundation for perceptual phenomena associated with Lis (2002) hypothesis that V1 builds a saliency map, and fits data on the tilt illusion.
****

A joint maximum-entropy model for binary neural population patterns and continuous signals
Sebastian Gerwinn, Philipp Berens, Matthias Bethge

Second-order maximum-entropy models have recently gained much interest for describing the statistics of binary spike trains. Here, we extend this approach to take continuous stimuli into account as well. By constraining the joint second-order statistics, we obtain a joint Gaussian-Boltzmann distribution of continuous stimuli and binary neural firing patterns, for which we also compute marginal and conditional distributions. This model has the same computational complexity as pure binary models and fitting it to data is a convex problem. We show that the model can be seen as an extension to the classical spike-triggered average/covariance analysis and can be used as a non-linear method for extracting features which a neural population is sensitive to. Further, by calculating the posterior distribution of stimuli given an observed neural response, the model can be used to decode stimuli and yields a natural spike-train metric. Therefore, extending the framework of maximum-entropy models to continuous variables allows us to gain novel insights into the relationship between the firing patterns of neural ensembles and the stimuli they are processing.
****

Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions
Matthias Bethge, Eero Simoncelli, Fabian Sinz
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.
****

Optimizing Multi-Class Spatio-Spectral Filters via Bayes Error Estimation for EEG Classification
Wenming Zheng, Zhouchen Lin
The method of common spatio-spectral patterns (CSSPs) is an extension of common spatial patterns (CSPs) by utilizing the technique of delay embedding to alleviate the adverse effects of noises and artifacts on the electroencephalogram (EEG) classification. Although the CSSPs method has shown to be more powerful than the CSPs method in the EEG classification, this method is only suitable for two-class EEG classification problems. In this paper, we generalize the two-class CSSPs method to multi-class cases. To this end, we first develop a novel theory of multi-class Bayes error estimation and then present the multi-class CSSPs (MCSSPs) method based on this Bayes error theoretical framework. By minimizing the estimated closed-form Bayes error, we obtain the optimal spatio-spectral filters of MCSSPs. To demonstrate the effectiveness of the proposed method, we conduct extensive experiments on the data set of BCI competition 2005. The experimental results show that our method significantly outperforms the previous multi-class CSPs (MCSPs) methods in the EEG classification.
****

Nash Equilibria of Static Prediction Games
Michael Brückner, Tobias Scheffer
The standard assumption of identically distributed training and test data can be violated when an adversary can exercise some control over the generation of the test data. In a prediction game, a learner produces a predictive model while an adversary may alter the distribution of input data. We study single-shot prediction games in which the cost functions of learner and adversary are not necessarily antagonistic. We identify conditions under which the prediction game has a unique Nash equilibrium, and derive algorithms that will find the equilibrial prediction models. In a case study, we explore properties of Nash-equilibrial prediction models for email spam filtering empirically.
****

Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization
John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, Yi Ma
Principal component analysis is a fundamental operation in computational data analysis, with myriad applications ranging from web search to bioinformatics to computer vision and image analysis. However, its performance and applicability in

real scenarios are limited by a lack of robustness to outlying or corrupted obse
rvations. This paper considers the idealized “robust principal component analysi
s” problem of recovering a low rank matrix A from corrupted observations D = A +
 E. Here, the error entries E can be arbitrarily large (modeling grossly corrupt
ed observations common in visual and bioinformatic data), but are assumed to be
sparse. We prove that most matrices A can be efficiently and exactly recovered f
rom most error sign-and-support patterns, by solving a simple convex program. Ou
r result holds even when the rank of A grows nearly proportionally (up to a loga
rithmic factor) to the dimensionality of the observation space and the number of
 errors E grows in proportion to the total number of entries in the matrix. A by
-product of our analysis is the first proportional growth results for the relate
d problem of completing a low-rank matrix from a small fraction of its entries.
Simulations and real-data examples corroborate the theoretical results, and sugg
est potential applications in computer vision.
****

Unsupervised Feature Selection for the k-means Clustering Problem
Christos Boutsidis, Petros Drineas, Michael W. Mahoney
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Multi-Step Dyna Planning for Policy Evaluation and Control
Hengshuai Yao, Shalabh Bhatnagar, Dongcui Diao, Richard S. Sutton, Csaba Szepesv
ári
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Variational Inference for the Nested Chinese Restaurant Process
Chong Wang, David Blei
The nested Chinese restaurant process (nCRP) is a powerful nonparametric Bayesia
n model for learning tree-based hierarchies from data. Since its posterior distr
ibution is intractable, current inference methods have all relied on MCMC sampli
ng. In this paper, we develop an alternative inference technique based on variat
ional methods. To employ variational methods, we derive a tree-based stick-break
ing construction of the nCRP mixture model, and a novel variational algorithm th
at efficiently explores a posterior over a large set of combinatorial structures
. We demonstrate the use of this approach for text and hand written digits model
ing, where we show we can adapt the nCRP to continuous data as well.
****

A Biologically Plausible Model for Rapid Natural Scene Identification
Sennay Ghebreab, Steven Scholte, Victor Lamme, Arnold Smeulders
Contrast statistics of the majority of natural images conform to a Weibull distr
ibution. This property of natural images may facilitate efficient and very rapid
 extraction of a scenes visual gist. Here we investigate whether a neural respon
se model based on the Weibull contrast distribution captures visual information
that humans use to rapidly identify natural scenes. In a learning phase, we meas
ure EEG activity of 32 subjects viewing brief flashes of 800 natural scenes. Fr
om these neural measurements and the contrast statistics of the natural image st
imuli, we derive an across subject Weibull response model. We use this model
to predict the responses to a large set of new scenes and estimate which scene t
he subject viewed by finding the best match between the model predictions and th
e observed EEG responses. In almost 90 percent of the cases our model accurately
 predicts the observed scene. Moreover, in most failed cases, the scene mistaken
 for the observed scene is visually similar to the observed scene itself. These
results suggest that Weibull contrast statistics of natural images contain a con
siderable amount of scene gist information to warrant rapid identification of na
tural images.

****
Learning from Neighboring Strokes: Combining Appearance and Context for Multi-Do
main Sketch Recognition
Tom Ouyang, Randall Davis
We propose a new sketch recognition framework that combines a rich representatio
n of low level visual appearance with a graphical model for capturing high level
 relationships between symbols. This joint model of appearance and context allow
s our framework to be less sensitive to noise and drawing variations, improving
accuracy and robustness. The result is a recognizer that is better able to handl
e the wide range of drawing styles found in messy freehand sketches. We evaluate
 our work on two real-world domains, molecular diagrams and electrical circuit d
iagrams, and show that our combined approach significantly improves recognition
performance.
****
Approximating MAP by Compensating for Structural Relaxations
Arthur Choi, Adnan Darwiche
We introduce a new perspective on approximations to the maximum a posteriori (MA
P) task in probabilistic graphical models, that is based on simplifying a given
instance, and then tightening the approximation. First, we start with a structu
ral relaxation of the original model. We then infer from the relaxation its def
iciencies, and compensate for them. This perspective allows us to identify two
distinct classes of approximations. First, we find that max-product belief prop
agation can be viewed as a way to compensate for a relaxation, based on a partic
ular idealized case for exactness. We identify a second approach to compensatio
n that is based on a more refined idealized case, resulting in a new approximati
on with distinct properties. We go on to propose a new class of algorithms that
, starting with a relaxation, iteratively yields tighter approximations.
****
Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations
Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro, Lawrence Carin, John Paisl
ey
Non-parametric Bayesian techniques are considered for learning dictionaries for
sparse image representations, with applications in denoising, inpainting and com
pressive sensing (CS). The beta process is employed as a prior for learning the
dictionary, and this non-parametric method naturally infers an appropriate dicti
onary size. The Dirichlet process and a probit stick-breaking process are also c
onsidered to exploit structure within an image. The proposed method can learn a
sparse dictionary in situ; training images may be exploited if available, but th
ey are not required. Further, the noise variance need not be known, and can be n
on-stationary. Another virtue of the proposed method is that sequential inferenc
e can be readily employed, thereby allowing scaling to large images. Several exa
mple results are presented, using both Gibbs and variational Bayesian inference,
 with comparisons to other state-of-the-art approaches.
****
Discrete MDL Predicts in Total Variation
Marcus Hutter
The Minimum Description Length (MDL) principle selects the model that has the sh
ortest code for data plus model. We show that for a countable class of models, M
DL predictions are close to the true distribution in a strong sense. The result
is completely general. No independence, ergodicity, stationarity, identifiabilit
y, or other assumption on the model class need to be made. More formally, we sho
w that for any countable class of models, the distributions selected by MDL (or
MAP) asymptotically predict (merge with) the true measure in the class in total
variation distance. Implications for non-i.i.d. domains like time-series forecas
ting, discriminative learning, and reinforcement learning are discussed.
****
Efficient and Accurate Lp-Norm Multiple Kernel Learning
Marius Kloft, Ulf Brefeld, Pavel Laskov, Klaus-Robert Müller, Alexander Zien, Sö
ren Sonnenburg
Learning linear combinations of multiple kernels is an appealing strategy when t

he right choice of features is unknown. Previous approaches to multiple kernel l
earning (MKL) promote sparse kernel combinations and hence support interpretabil
ity. Unfortunately, L1-norm MKL is hardly observed to outperform trivial baselin
es in practical applications. To allow for robust kernel mixtures, we generalize
 MKL to arbitrary Lp-norms. We devise new insights on the connection between se
veral existing MKL formulations and develop two efficient interleaved optimizati
on strategies for arbitrary p>1. Empirically, we demonstrate that the interleave
d optimization strategies are much faster compared to the traditionally used wra
pper approaches. Finally, we apply Lp-norm MKL to real-world problems from compu
tational biology, showing that non-sparse MKL achieves accuracies that go beyond
 the state-of-the-art.
****

Occlusive Components Analysis
Jörg Lücke, Richard Turner, Maneesh Sahani, Marc Henniges
We study unsupervised learning in a probabilistic generative model for occlusion
. The model uses two types of latent variables: one indicates which objects are
present in the image, and the other how they are ordered in depth. This depth or
der then determines how the positions and appearances of the objects present, sp
ecified in the model parameters, combine to form the image. We show that the obj
ect parameters can be learnt from an unlabelled set of images in which objects o
cclude one another. Exact maximum-likelihood learning is intractable. However, w
e show that tractable approximations to Expectation Maximization (EM) can be fou
nd if the training images each contain only a small number of objects on average
. In numerical experiments it is shown that these approximations recover the cor
rect set of object parameters. Experiments on a novel version of the bars test u
sing colored bars, and experiments on more realistic data, show that the algorit
hm performs well in extracting the generating causes. Experiments based on the s
tandard bars benchmark test for object learning show that the algorithm performs
 well in comparison to other recent component extraction approaches. The model a
nd the learning algorithm thus connect research on occlusion with the research f
ield of multiple-cause component extraction methods.
****

Distribution-Calibrated Hierarchical Classification
Ofer Dekel
While many advances have already been made on the topic of hierarchical classi-
 ■cation learning, we take a step back and examine how a hierarchical classi■ca-
  tion problem should be formally de■ned. We pay particular attention to the fac
t that many arbitrary decisions go into the design of the the label taxonomy th
at is provided with the training data, and that this taxonomy is often unbalanc
ed. We correct this problem by using the data distribution to calibrate the hie
rarchical classi■cation loss function. This distribution-based correction must
be done with care, to avoid introducing unmanagable statstical dependencies int
o the learning problem. This leads us off the beaten path of binomial-type esti
mation and into the uncharted waters of geometric-type estimation. We present a
 new calibrated de■nition of statistical risk for hierarchical classi■cation, a
n unbiased geometric estimator for this risk, and a new algorithmic reduction f
rom hierarchical classi■- cation to cost-sensitive classi■cation.
****

A Smoothed Approximate Linear Program
Vijay Desai, Vivek Farias, Ciamac C. Moallemi
We present a novel linear program for the approximation of the dynamic programmi
ng cost-to-go function in high-dimensional stochastic control problems. LP appro
aches to approximate DP naturally restrict attention to approximations that are
lower bounds to the optimal cost-to-go function. Our program -- the `smoothed ap
proximate linear program -- relaxes this restriction in an appropriate fashion w
hile remaining computationally tractable. Doing so appears to have several advan
tages: First, we demonstrate superior bounds on the quality of approximation to
the optimal cost-to-go function afforded by our approach. Second, experiments wi
th our approach on a challenging problem (the game of Tetris) show that the appr
oach outperforms the existing LP approach (which has previously been shown to be

competitive with several ADP algorithms) by an order of magnitude.
****
Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model
Ed Vul, George Alvarez, Joshua Tenenbaum, Michael Black
Multiple object tracking is a task commonly used to investigate the architecture of human visual attention. Human participants show a distinctive pattern of successes and failures in tracking experiments that is often attributed to limits on an object system, a tracking module, or other specialized cognitive structures. Here we use a computational analysis of the task of object tracking to ask which human failures arise from cognitive limitations and which are consequences of inevitable perceptual uncertainty in the tracking task. We find that many human performance phenomena, measured through novel behavioral experiments, are naturally produced by the operation of our ideal observer model (a Rao-Blackwelized particle filter). The tradeoff between the speed and number of objects being tracked, however, can only arise from the allocation of a flexible cognitive resource, which can be formalized as either memory or attention.
****
Graph-based Consensus Maximization among Multiple Supervised and Unsupervised Models
Jing Gao, Feng Liang, Wei Fan, Yizhou Sun, Jiawei Han
Little work has been done to directly combine the outputs of multiple supervised and unsupervised models. However, it can increase the accuracy and applicability of ensemble methods. First, we can boost the diversity of classification ensemble by incorporating multiple clustering outputs, each of which provides grouping constraints for the joint label predictions of a set of related objects. Secondly, ensemble of supervised models is limited in applications which have no access to raw data but to the meta-level model outputs. In this paper, we aim at calculating a consolidated classification solution for a set of objects by maximizing the consensus among both supervised predictions and unsupervised grouping constraints. We seek a global optimal label assignment for the target objects, which is different from the result of traditional majority voting and model combination approaches. We cast the problem into an optimization problem on a bipartite graph, where the objective function favors smoothness in the conditional probability estimates over the graph, as well as penalizes deviation from initial labeling of supervised models. We solve the problem through iterative propagation of conditional probability estimates among neighboring nodes, and interpret the method as conducting a constrained embedding in a transformed space, as well as a ranking on the graph. Experimental results on three real applications demonstrate the benefits of the proposed method over existing alternatives.
****
Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models
Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, Gideon Mann
Training conditional maximum entropy models on massive data requires significant time and computational resources. In this paper, we investigate three common distributed training strategies: distributed gradient, majority voting ensembles, and parameter mixtures. We analyze the worst-case runtime and resource costs of each and present a theoretical foundation for the convergence of parameters under parameter mixtures, the most efficient strategy. We present large-scale experiments comparing the different strategies and demonstrate that parameter mixtures over independent models use fewer resources and achieve comparable loss as compared to standard approaches.
****
On the Algorithmics and Applications of a Mixed-norm based Kernel Learning Formulation
Saketha Jagarlapudi, Dinesh G, Raman S, Chiranjib Bhattacharyya, Aharon Ben-tal, Ramakrishnan K.r.
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****

Estimating image bases for visual image reconstruction from human brain activity
Yusuke Fujiwara, Yoichi Miyawaki, Yukiyasu Kamitani
Image representation based on image bases provides a framework for understanding neural representation of visual perception. A recent fMRI study has shown that arbitrary contrast-defined visual images can be reconstructed from fMRI activity patterns using a combination of multi-scale local image bases. In the reconstruction model, the mapping from an fMRI activity pattern to the contrasts of the image bases was learned from measured fMRI responses to visual images. But the shapes of the images bases were fixed, and thus may not be optimal for reconstruction. Here, we propose a method to build a reconstruction model in which image bases are automatically extracted from the measured data. We constructed a probabilistic model that relates the fMRI activity space to the visual image space via a set of latent variables. The mapping from the latent variables to the visual image space can be regarded as a set of image bases. We found that spatially localized, multi-scale image bases were estimated near the fovea, and that the model using the estimated image bases was able to accurately reconstruct novel visual images. The proposed method provides a means to discover a novel functional mapping between stimuli and brain activity patterns.
****

A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers
Sahand Negahban, Bin Yu, Martin J. Wainwright, Pradeep Ravikumar
The estimation of high-dimensional parametric models requires imposing some structure on the models, for instance that they be sparse, or that matrix structured parameters have low rank. A general approach for such structured parametric model estimation is to use regularized M-estimation procedures, which regularize a loss function that measures goodness of fit of the parameters to the data with some regularization function that encourages the assumed structure. In this paper, we aim to provide a unified analysis of such regularized M-estimation procedures. In particular, we report the convergence rates of such estimators in any metric norm. Using just our main theorem, we are able to rederive some of the many existing results, but also obtain a wide range of novel convergence rates results. Our analysis also identifies key properties of loss and regularization functions such as restricted strong convexity, and decomposability, that ensure the corresponding regularized M-estimators have good convergence rates.
****

Grouped Orthogonal Matching Pursuit for Variable Selection and Prediction
Grzegorz Swirszcz, Naoki Abe, Aurelie C. Lozano
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.
****

Rank-Approximate Nearest Neighbor Search: Retaining Meaning and Speed in High Dimensions
Parikshit Ram, Dongryeol Lee, Hua Ouyang, Alexander Gray
The long-standing problem of efficient nearest-neighbor (NN) search has ubiquitous applications ranging from astrophysics to MP3 fingerprinting to bioinformatics to movie recommendations. As the dimensionality of the dataset increases, exact NN search becomes computationally prohibitive; (1+eps)-distance-approximate NN search can provide large speedups but risks losing the meaning of NN search present in the ranks (ordering) of the distances. This paper presents a simple, practical algorithm allowing the user to, for the first time, directly control the true accuracy of NN search (in terms of ranks) while still achieving the large speedups over exact NN. Experiments with high-dimensional datasets show that it often achieves faster and more accurate results than the best-known distance-approximate method, with much more stable behavior.
****

Online Learning of Assignments
Matthew Streeter, Daniel Golovin, Andreas Krause
Which ads should we display in sponsored search in order to maximize our revenue ? How should we dynamically rank information sources to maximize value of information? These applications exhibit strong diminishing returns: Selection of redundant ads and information sources decreases their marginal utility. We show that these and other problems can be formalized as repeatedly selecting an assignment of items to positions to maximize a sequence of monotone submodular functions that arrive one by one. We present an efficient algorithm for this general problem and analyze it in the no-regret model. Our algorithm is equipped with strong theoretical guarantees, with a performance ratio that converges to the optimal constant of 1-1/e. We empirically evaluate our algorithms on two real-world online optimization problems on the web: ad allocation with submodular utilities, and dynamically ranking blogs to detect information cascades.
****

Skill Discovery in Continuous Reinforcement Learning Domains using Skill Chaining
George Konidaris, Andrew Barto
We introduce skill chaining, a skill discovery method for reinforcement learning agents in continuous domains, that builds chains of skills leading to an end-of-task reward. We demonstrate experimentally that it creates skills that result in performance benefits in a challenging continuous domain.
****

Strategy Grafting in Extensive Games
Kevin Waugh, Nolan Bard, Michael Bowling
Extensive games are often used to model the interactions of multiple agents within an environment. Much recent work has focused on increasing the size of an extensive game that can be feasibly solved. Despite these improvements, many interesting games are still too large for such techniques. A common approach for computing strategies in these large games is to first employ an abstraction technique to reduce the original game to an abstract game that is of a manageable size. This abstract game is then solved and the resulting strategy is used in the original game. Most top programs in recent AAAI Computer Poker Competitions use this approach. The trend in this competition has been that strategies found in larger abstract games tend to beat strategies found in smaller abstract games. These larger abstract games have more expressive strategy spaces and therefore contain better strategies. In this paper we present a new method for computing strategies in large games. This method allows us to compute more expressive strategies without increasing the size of abstract games that we are required to solve. We demonstrate the power of the approach experimentally in both small and large games, while also providing a theoretical justification for the resulting improvement.
****

Training Factor Graphs with Reinforcement Learning for Efficient MAP Inference
Khashayar Rohanimanesh, Sameer Singh, Andrew McCallum, Michael Black
Large, relational factor graphs with structure defined by first-order logic or other languages give rise to notoriously difficult inference problems. Because unrolling the structure necessary to represent distributions over all hypotheses has exponential blow-up, solutions are often derived from MCMC. However, because of limitations in the design and parameterization of the jump function, these sampling-based methods suffer from local minima|the system must transition through lower-scoring configurations before arriving at a better MAP solution. This paper presents a new method of explicitly selecting fruitful downward jumps by leveraging reinforcement learning (RL). Rather than setting parameters to maximize the likelihood of the training data, parameters of the factor graph are treated as a log-linear function approximator and learned with temporal difference (TD); MAP inference is performed by executing the resulting policy on held out test data. Our method allows efficient gradient updates since only factors in the neighborhood of variables affected by an action need to be computed|we bypass the need to compute marginals entirely. Our method

provides dramatic empirical success, producing new state-of-the-art results o
n a complex joint model of ontology alignment, with a 48\% reduction in error
over state-of-the-art in that domain.
****

Learning a Small Mixture of Trees
M. Kumar, Daphne Koller
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

An Additive Latent Feature Model for Transparent Object Recognition
Mario Fritz, Gary Bradski, Sergey Karayev, Trevor Darrell, Michael Black
Existing methods for recognition of object instances and categories based on qua
ntized local features can perform poorly when local features exist on transparen
t surfaces, such as glass or plastic objects. There are characteristic patterns
to the local appearance of transparent objects, but they may not be well capture
d by distances to individual examples or by a local pattern codebook obtained by
vector quantization. The appearance of a transparent patch is determined in pa
rt by the refraction of a background pattern through a transparent medium: the e
nergy from the background usually dominates the patch appearance. We model tran
sparent local patch appearance using an additive model of latent factors: backg
round factors due to scene content, and factors which capture a local edge ener
gy distribution characteristic of the refraction. We implement our method using
a novel LDA-SIFT formulation which performs LDA prior to any vector quantizatio
n step; we discover latent topics which are characteristic of particular transpa
rent patches and quantize the SIFT space into transparent visual words according
to the latent topic dimensions. No knowledge of the background scene is require
d at test time; we show examples recognizing transparent glasses in a domestic e
nvironment.
****

Know Thy Neighbour: A Normative Theory of Synaptic Depression
Jean-pascal Pfister, Peter Dayan, Máté Lengyel
Synapses exhibit an extraordinary degree of short-term malleability, with releas
e probabilities and effective synaptic strengths changing markedly over multiple
timescales. From the perspective of a fixed computational operation in a networ
k, this seems like a most unacceptable degree of added noise. We suggest an alte
rnative theory according to which short term synaptic plasticity plays a normati
vely-justifiable role. This theory starts from the commonplace observation that
the spiking of a neuron is an incomplete, digital, report of the analog quantity
that contains all the critical information, namely its membrane potential. We s
uggest that one key task for a synapse is to solve the inverse problem of estima
ting the pre-synaptic membrane potential from the spikes it receives and prior
expectations, as in a recursive filter. We show that short-term synaptic depress
ion has canonical dynamics which closely resemble those required for optimal est
imation, and that it indeed supports high quality estimation. Under this account
, the local postsynaptic potential and the level of synaptic resources track the
(scaled) mean and variance of the estimated presynaptic membrane potential. We
make experimentally testable predictions for how the statistics of subthreshold
membrane potential fluctuations and the form of spiking non-linearity should be
related to the properties of short-term plasticity in any particular cell type.
****

Randomized Pruning: Efficiently Calculating Expectations in Large Dynamic Progra
ms
Alexandre Bouchard-côté, Slav Petrov, Dan Klein
Pruning can massively accelerate the computation of feature expectations in larg
e models. However, any single pruning mask will introduce bias. We present a n
ovel approach which employs a randomized sequence of pruning masks. Formally, we
apply auxiliary variable MCMC sampling to generate this sequence of masks, ther
eby gaining theoretical guarantees about convergence. Because each mask is gener

ally able to skip large portions of an underlying dynamic program, our approach is particularly compelling for high-degree algorithms. Empirically, we demonstrate our method on bilingual parsing, showing decreasing bias as more masks are incorporated, and outperforming fixed tic-tac-toe pruning.
****

A Rate Distortion Approach for Semi-Supervised Conditional Random Fields
Yang Wang, Gholamreza Haffari, Shaojun Wang, Greg Mori
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.
****

Learning to Explore and Exploit in POMDPs
Chenghui Cai, Xuejun Liao, Lawrence Carin
A fundamental objective in reinforcement learning is the maintenance of a proper balance between exploration and exploitation. This problem becomes more challenging when the agent can only partially observe the states of its environment. In this paper we propose a dual-policy method for jointly learning the agent behavior and the balance between exploration exploitation, in partially observable environments. The method subsumes traditional exploration, in which the agent takes actions to gather information about the environment, and active learning, in which the agent queries an oracle for optimal actions (with an associated cost for employing the oracle). The form of the employed exploration is dictated by the specific problem. Theoretical guarantees are provided concerning the optimality of the balancing of exploration and exploitation. The effectiveness of the method is demonstrated by experimental results on benchmark problems.
****

Multiple Incremental Decremental Learning of Support Vector Machines
Masayuki Karasuyama, Ichiro Takeuchi
We propose a multiple incremental decremental algorithm of Support Vector Machine (SVM). Conventional single cremental decremental SVM can update the trained model efficiently when single data point is added to or removed from the training set. When we add and/or remove multiple data points, this algorithm is time-consuming because we need to repeatedly apply it to each data point. The roposed algorithm is computationally more efficient when multiple data points are added and/or removed simultaneously. The single incremental decremental algorithm is built on an optimization technique called parametric programming. We extend the idea and introduce multi-parametric programming for developing the proposed algorithm. Experimental results on synthetic and real data sets indicate that the proposed algorithm can significantly reduce the computational cost of multiple incremental decremental operation. Our approach is especially useful for online SVM learning in which we need to remove old data points and add new data points in a short amount of time.
****

Bayesian Source Localization with the Multivariate Laplace Prior
Marcel Gerven, Botond Cseke, Robert Oostenveld, Tom Heskes
We introduce a novel multivariate Laplace (MVL) distribution as a sparsity promoting prior for Bayesian source localization that allows the specification of constraints between and within sources. We represent the MVL distribution as a scale mixture that induces a coupling between source variances instead of their means. Approximation of the posterior marginals using expectation propagation is shown to be very efficient due to properties of the scale mixture representation. The computational bottleneck amounts to computing the diagonal elements of a sparse matrix inverse. Our approach is illustrated using a mismatch negativity paradigm for which MEG data and a structural MRI have been acquired. We show that spatial coupling leads to sources which are active over larger cortical areas as compared with an uncoupled prior.
****

Learning Non-Linear Combinations of Kernels
Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh

This paper studies the general problem of learning kernels based on a polynomial combination of base kernels. It analyzes this problem in the case of regression and the kernel ridge regression algorithm. It examines the corresponding learning kernel optimization problem, shows how that minimax problem can be reduced to a simpler minimization problem, and proves that the global solution of this problem always lies on the boundary. We give a projection-based gradient descent algorithm for solving the optimization problem, shown empirically to converge in few iterations. Finally, we report the results of extensive experiments with this algorithm using several publicly available datasets demonstrating the effectiveness of our technique.

****

Conditional Neural Fields

Jian Peng, Liefeng Bo, Jinbo Xu

Conditional random fields (CRF) are quite successful on sequence labeling tasks such as natural language processing and biological sequence analysis. CRF models use linear potential functions to represent the relationship between input features and outputs. However, in many real-world applications such as protein structure prediction and handwriting recognition, the relationship between input features and outputs is highly complex and nonlinear, which cannot be accurately modeled by a linear function. To model the nonlinear relationship between input features and outputs we propose Conditional Neural Fields (CNF), a new conditional probabilistic graphical model for sequence labeling. Our CNF model extends CRF by adding one (or possibly several) middle layer between input features and outputs. The middle layer consists of a number of hidden parameterized gates, each acting as a local neural network node or feature extractor to capture the nonlinear relationship between input features and outputs. Therefore, conceptually this CNF model is much more expressive than the linear CRF model. To better control the complexity of the CNF model, we also present a hyperparameter optimization procedure within the evidence framework. Experiments on two widely-used benchmarks indicate that this CNF model performs significantly better than a number of popular methods. In particular, our CNF model is the best among about ten machine learning methods for protein secondary tructure prediction and also among a few of the best methods for handwriting recognition.

****

An Online Algorithm for Large Scale Image Similarity Learning

Gal Chechik, Uri Shalit, Varun Sharma, Samy Bengio

Learning a measure of similarity between pairs of objects is a fundamental problem in machine learning. It stands in the core of classification methods like kernel machines, and is particularly useful for applications like searching for images that are similar to a given image or finding videos that are relevant to a given video. In these tasks, users look for objects that are not only visually similar but also semantically related to a given object. Unfortunately, current approaches for learning similarity may not scale to large datasets with high dimensionality, especially when imposing metric constraints on the learned similarity. We describe OASIS, a method for learning pairwise similarity that is fast and scales linearly with the number of objects and the number of non-zero features. Scalability is achieved through online learning of a bilinear model over sparse representations using a large margin criterion and an efficient hinge loss cost. OASIS is accurate at a wide range of scales: on a standard benchmark with thousands of images, it is more precise than state-of-the-art methods, and faster by orders of magnitude. On 2 million images collected from the web, OASIS can be trained within 3 days on a single CPU. The non-metric similarities learned by OASIS can be transformed into metric similarities, achieving higher precisions than similarities that are learned as metrics in the first place. This suggests an approach for learning a metric from data that is larger by an order of magnitude than was handled before.

****

Functional network reorganization in motor cortex can be explained by reward-modulated Hebbian learning

Steven Chase, Andrew Schwartz, Wolfgang Maass, Robert Legenstein

The control of neuroprosthetic devices from the activity of motor cortex neurons benefits from learning effects where the function of these neurons is adapted to the control task. It was recently shown that tuning properties of neurons in monkey motor cortex are adapted selectively in order to compensate for an erroneous interpretation of their activity. In particular, it was shown that the tuning curves of those neurons whose preferred directions had been misinterpreted changed more than those of other neurons. In this article, we show that the experimentally observed self-tuning properties of the system can be explained on the basis of a simple learning rule. This learning rule utilizes neuronal noise for exploration and performs Hebbian weight updates that are modulated by a global reward signal. In contrast to most previously proposed reward-modulated Hebbian learning rules, this rule does not require extraneous knowledge about what is noise and what is signal. The learning rule is able to optimize the performance of the model system within biologically realistic periods of time and under high noise levels. When the neuronal noise is fitted to experimental data, the model produces learning effects similar to those found in monkey experiments.

****

The Infinite Partially Observable Markov Decision Process

Finale Doshi-velez

The Partially Observable Markov Decision Process (POMDP) framework has proven useful in planning domains that require balancing actions that increase an agents knowledge and actions that increase an agents reward. Unfortunately, most POMDPs are complex structures with a large number of parameters. In many realworld problems, both the structure and the parameters are difficult to specify from domain knowledge alone. Recent work in Bayesian reinforcement learning has made headway in learning POMDP models; however, this work has largely focused on learning the parameters of the POMDP model. We define an infinite POMDP (iPOMDP) model that does not require knowledge of the size of the state space; instead, it assumes that the number of visited states will grow as the agent explores its world and explicitly models only visited states. We demonstrate the iPOMDP utility on several standard problems.

****

Accelerated Gradient Methods for Stochastic Optimization and Online Learning

Chonghai Hu, Weike Pan, James Kwok

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.

****

Robust Value Function Approximation Using Bilinear Programming

Marek Petrik, Shlomo Zilberstein

Existing value function approximation methods have been successfully used in many applications, but they often lack useful a priori error bounds. We propose approximate bilinear programming, a new formulation of value function approximation that provides strong a priori guarantees. In particular, it provably finds an approximate value function that minimizes the Bellman residual. Solving a bilinear program optimally is NP hard, but this is unavoidable because the Bellman-residual minimization itself is NP hard. We, therefore, employ and analyze a common approximate algorithm for bilinear programs. The analysis shows that this algorithm offers a convergent generalization of approximate policy iteration. Finally, we demonstrate that the proposed approach can consistently minimize the Bellman residual on a simple benchmark problem.

****

Parallel Inference for Latent Dirichlet Allocation on Graphics Processing Units

Feng Yan, Ningyi Xu, Yuan Qi

The recent emergence of Graphics Processing Units (GPUs) as general-purpose parallel computing devices provides us with new opportunities to develop scalable learning methods for massive data. In this work, we consider the problem of parallelizing two inference methods on GPUs for latent Dirichlet Allocation (LDA) models, collapsed Gibbs sampling (CGS) and collapsed variational Bayesian (CVB). To

address limited memory constraints on GPUs, we propose a novel data partitioning scheme that effectively reduces the memory cost. Furthermore, the partitioning scheme balances the computational cost on each multiprocessor and enables us to easily avoid memory access conflicts. We also use data streaming to handle extremely large datasets. Extensive experiments showed that our parallel inference methods consistently produced LDA models with the same predictive power as sequential training methods did but with 26x speedup for CGS and 196x speedup for CVB on a GPU with 30 multiprocessors; actually the speedup is almost linearly scalable with the number of multiprocessors available. The proposed partitioning scheme and data streaming can be easily ported to many other models in machine learning.

****

Local Rules for Global MAP: When Do They Work ?

Kyomin Jung, Pushmeet Kohli, 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.

****

Bayesian estimation of orientation preference maps

Sebastian Gerwinn, Leonard White, Matthias Kaschube, Matthias Bethge, Jakob H. Macke

Imaging techniques such as optical imaging of intrinsic signals, 2-photon calcium imaging and voltage sensitive dye imaging can be used to measure the functional organization of visual cortex across different spatial scales. Here, we present Bayesian methods based on Gaussian processes for extracting topographic maps from functional imaging data. In particular, we focus on the estimation of orientation preference maps (OPMs) from intrinsic signal imaging data. We model the underlying map as a bivariate Gaussian process, with a prior covariance function that reflects known properties of OPMs, and a noise covariance adjusted to the data. The posterior mean can be interpreted as an optimally smoothed estimate of the map, and can be used for model based interpolations of the map from sparse measurements. By sampling from the posterior distribution, we can get error bars on statistical properties such as preferred orientations, pinwheel locations or -counts. Finally, the use of an explicit probabilistic model facilitates interpretation of parameters and provides the basis for decoding studies. We demonstrate our model both on simulated data and on intrinsic signaling data from ferret visual cortex.

****

A Bayesian Analysis of Dynamics in Free Recall

Richard Socher, Samuel Gershman, Per Sederberg, Kenneth Norman, Adler Perotte, David Blei

We develop a probabilistic model of human memory performance in free recall experiments. In these experiments, a subject first studies a list of words and then tries to recall them. To model these data, we draw on both previous psychological research and statistical topic models of text documents. We assume that memories are formed by assimilating the semantic meaning of studied words (represented as a distribution over topics) into a slowly changing latent context (represented in the same space). During recall, this context is reinstated and used as a cue for retrieving studied words. By conceptualizing memory retrieval as a dynamic latent variable model, we are able to use Bayesian inference to represent uncertainty and reason about the cognitive processes underlying memory. We present a particle filter algorithm for performing approximate posterior inference, and evaluate our model on the prediction of recalled words in experimental data. By specifying the model hierarchically, we are also able to capture inter-subject variability.

****

Structural inference affects depth perception in the context of potential occlusion

Ian Stevenson, Konrad Koerding

In many domains, humans appear to combine perceptual cues in a near-optimal, pro babilistic fashion: two noisy pieces of information tend to be combined linearly with weights proportional to the precision of each cue. Here we present a case where structural information plays an important role. The presence of a backgrou nd cue gives rise to the possibility of occlusion, and places a soft constraint on the location of a target – in effect propelling it forward. We present an ide al observer model of depth estimation for this situation where structural or ord inal information is important and then fit the model to human data from a stereo -matching task. To test whether subjects are truly using ordinal cues in a proba bilistic manner we then vary the uncertainty of the task. We find that the model accurately predicts shifts in subject's behavior. Our results indicate that the nervous system estimates depth ordering in a probabilistic fashion and estimate s the structure of the visual scene during depth perception.

****

Indian Buffet Processes with Power-law Behavior

Yee Teh, Dilan Gorur

The Indian buffet process (IBP) is an exchangeable distribution over binary matr ices used in Bayesian nonparametric featural models. In this paper we propose a three-parameter generalization of the IBP exhibiting power-law behavior. We ac hieve this by generalizing the beta process (the de Finetti measure of the IBP) to the \emph{stable-beta process} and deriving the IBP corresponding to it. We find interesting relationships between the stable-beta process and the Pitman-Yo r process (another stochastic process used in Bayesian nonparametric models with interesting power-law properties). We show that our power-law IBP is a good mo del for word occurrences in documents with improved performance over the normal IBP.

****

Boosting with Spatial Regularization

Yongxin Xi, Uri Hasson, Peter J. Ramadge, Zhen Xiang

By adding a spatial regularization kernel to a standard loss function formulatio n of the boosting problem, we develop a framework for spatially informed boostin g. From this regularized loss framework we derive an efficient boosting algorit hm that uses additional weights/priors on the base classifiers. We prove that t he proposed algorithm exhibits a ``grouping effect, which encourages the selecti on of all spatially local, discriminative base classifiers. The algorithms prim ary advantage is in applications where the trained classifier is used to identif y the spatial pattern of discriminative information, e.g. the voxel selection pr oblem in fMRI. We demonstrate the algorithms performance on various data sets.

****

Time-rescaling methods for the estimation and assessment of non-Poisson neural e ncoding models

Jonathan Pillow

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

****

Semi-supervised Regression using Hessian energy with an application to semi-supe rvised dimensionality reduction

Kwang Kim, Florian Steinke, Matthias Hein

Semi-supervised regression based on the graph Laplacian suffers from the fact th at the solution is biased towards a constant and the lack of extrapolating power . Outgoing from these observations we propose to use the second-order Hessian en ergy for semi-supervised regression which overcomes both of these problems, in p articular, if the data lies on or close to a low-dimensional submanifold in the feature space, the Hessian energy prefers functions which vary ``linearly with r espect to the natural parameters in the data. This property makes it also partic ularly suited for the task of semi-supervised dimensionality reduction where the goal is to find the natural parameters in the data based on a few labeled point s. The experimental result suggest that our method is superior to semi-supervise

d regression using Laplacian regularization and standard supervised methods and is particularly suited for semi-supervised dimensionality reduction.
****

Localizing Bugs in Program Executions with Graphical Models
Laura Dietz, Valentin Dallmeier, Andreas Zeller, Tobias Scheffer
We devise a graphical model that supports the process of debugging software by g uiding developers to code that is likely to contain defects. The model is traine d using execution traces of passing test runs; it reflects the distribution over transitional patterns of code positions. Given a failing test case, the model d etermines the least likely transitional pattern in the execution trace. The mode l is designed such that Bayesian inference has a closed-form solution. We evalua te the Bernoulli graph model on data of the software projects AspectJ and Rhino .
****

Human Rademacher Complexity
Jerry Zhu, Bryan Gibson, Timothy T. Rogers
We propose to use Rademacher complexity, originally developed in computational l earning theory, as a measure of human learning capacity. Rademacher complexity measures a learners ability to fit random data, and can be used to bound the lea rners true error based on the observed training sample error. We first review t he definition of Rademacher complexity and its generalization bound. We then de scribe a learning the noise" procedure to experimentally measure human Rademache r complexities. The results from empirical studies showed that: (i) human Radem acher complexity can be successfully measured, (ii) the complexity depends on th e domain and training sample size in intuitive ways, (iii) human learning respec ts the generalization bounds, (iv) the bounds can be useful in predicting the da nger of overfitting in human learning. Finally, we discuss the potential applic ations of human Rademacher complexity in cognitive science."
****

Beyond Convexity: Online Submodular Minimization
Elad Hazan, Satyen Kale
We consider an online decision problem over a discrete space in which the loss f unction is submodular. We give algorithms which are computationally efficient an d are Hannan-consistent in both the full information and bandit settings.
****

Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization
Massih R. Amini, Nicolas Usunier, Cyril Goutte
We address the problem of learning classifiers when observations have multiple v iews, some of which may not be observed for all examples. We assume the existen ce of view generating functions which may complete the missing views in an appro ximate way. This situation corresponds for example to learning text classifiers from multilingual collections where documents are not available in all language s. In that case, Machine Translation (MT) systems may be used to translate each document in the missing languages. We derive a generalization error bound for classifiers learned on examples with multiple artificially created views. Our re sult uncovers a trade-off between the size of the training set, the number of vi ews, and the quality of the view generating functions. As a consequence, we iden tify situations where it is more interesting to use multiple views for learning instead of classical single view learning. An extension of this framework is a natural way to leverage unlabeled multi-view data in semi-supervised learning. Experimental results on a subset of the Reuters RCV1/RCV2 collections support ou r findings by showing that additional views obtained from MT may significantly i mprove the classification performance in the cases identified by our trade-off.
****

Measuring model complexity with the prior predictive
Wolf Vanpaemel
In the last few decades, model complexity has received a lot of press. While man y methods have been proposed that jointly measure a model’s descriptive adequacy and its complexity, few measures exist that measure complexity in itself. Moreo

ver, existing measures ignore the parameter prior, which is an inherent part of the model and affects the complexity. This paper presents a stand alone measure for model complexity, that takes the number of parameters, the functional form, the range of the parameters and the parameter prior into account. This Prior Predictive Complexity (PPC) is an intuitive and easy to compute measure. It starts from the observation that model complexity is the property of the model that enables it to fit a wide range of outcomes. The PPC then measures how wide this range exactly is.
****

Nonparametric Bayesian Texture Learning and Synthesis
Long Zhu, Yuanahao Chen, Bill Freeman, Antonio Torralba
We present a nonparametric Bayesian method for texture learning and synthesis. A texture image is represented by a 2D-Hidden Markov Model (2D-HMM) where the hidden states correspond to the cluster labeling of textons and the transition matrix encodes their spatial layout (the compatibility between adjacent textons). 2D-HMM is coupled with the Hierarchical Dirichlet process (HDP) which allows the number of textons and the complexity of transition matrix grow as the input texture becomes irregular. The HDP makes use of Dirichlet process prior which favors regular textures by penalizing the model complexity. This framework (HDP-2D-HMM) learns the texton vocabulary and their spatial layout jointly and automatically. The HDP-2D-HMM results in a compact representation of textures which allows fast texture synthesis with comparable rendering quality over the state-of-the-art image-based rendering methods. We also show that HDP-2D-HMM can be applied to perform image segmentation and synthesis.
****

A Bayesian Model for Simultaneous Image Clustering, Annotation and Object Segmentation
Lan Du, Lu Ren, Lawrence Carin, David Dunson
A non-parametric Bayesian model is proposed for processing multiple images. The analysis employs image features and, when present, the words associated with accompanying annotations. The model clusters the images into classes, and each image is segmented into a set of objects, also allowing the opportunity to assign a word to each object (localized labeling). Each object is assumed to be represented as a heterogeneous mix of components, with this realized via mixture models linking image features to object types. The number of image classes, number of object types, and the characteristics of the object-feature mixture models are inferred non-parametrically. To constitute spatially contiguous objects, a new logistic stick-breaking process is developed. Inference is performed efficiently via variational Bayesian analysis, with example results presented on two image databases.
****

Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise
Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, Paul Ruvolo
Modern machine learning-based approaches to computer vision require very large databases of labeled images. Some contemporary vision systems already require on the order of millions of images for training (e.g., Omron face detector). While the collection of these large databases is becoming a bottleneck, new Internet-based services that allow labelers from around the world to be easily hired and managed provide a promising solution. However, using these services to label large databases brings with it new theoretical and practical challenges: (1) The labelers may have wide ranging levels of expertise which are unknown a priori, and in some cases may be adversarial; (2) images may vary in their level of difficulty; and (3) multiple labels for the same image must be combined to provide an estimate of the actual label of the image. Probabilistic approaches provide a principled way to approach these problems. In this paper we present a probabilistic model and use it to simultaneously infer the label of each image, the expertise of each labeler, and the difficulty of each image. On both simulated and real data, we demonstrate that the model outperforms the commonly used ``Majority Vote heuristic for inferring image labels, and is robust to both adversarial and noi

sy labelers.
****

Reading Tea Leaves: How Humans Interpret Topic Models
Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, David Blei
Probabilistic topic models are a popular tool for the unsupervised analysis of t ext, providing both a predictive model of future text and a latent topic represe ntation of the corpus. Practitioners typically assume that the latent space is s emantically meaningful. It is used to check models, summarize the corpus, and gu ide exploration of its contents. However, whether the latent space is interpreta ble is in need of quantitative evaluation. In this paper, we present new quantit ative methods for measuring semantic meaning in inferred topics. We back these m easures with large-scale user studies, showing that they capture aspects of the model that are undetected by previous measures of model quality based on held-ou t likelihood. Surprisingly, topic models which perform better on held-out likeli hood may infer less semantically meaningful topics.
****

Segmenting Scenes by Matching Image Composites
Bryan Russell, Alyosha Efros, Josef Sivic, Bill Freeman, Andrew Zisserman
In this paper, we investigate how similar images sharing the same global descrip tion can help with unsupervised scene segmentation in an image. In contrast to recent work in semantic alignment of scenes, we allow an input image to be expla ined by partial matches of similar scenes. This allows for a better explanation of the input scenes. We perform MRF-based segmentation that optimizes over mat ches, while respecting boundary information. The recovered segments are then us ed to re-query a large database of images to retrieve better matches for the tar get region. We show improved performance in detecting occluding boundaries over previous methods on data gathered from the LabelMe database.
****

Generalization Errors and Learning Curves for Regression with Multi-task Gaussia n Processes
Kian Chai
We provide some insights into how task correlations in multi-task Gaussian proce ss (GP) regression affect the generalization error and the learning curve. We a nalyze the asymmetric two-task case, where a secondary task is to help the learn ing of a primary task. Within this setting, we give bounds on the generalization error and the learning curve of the primary task. Our approach admits intuitive understandings of the multi-task GP by relating it to single-task GPs. For the case of one-dimensional input-space under optimal sampling with data only for th e secondary task, the limitations of multi-task GP can be quantified explicitly.
****

Breaking Boundaries Between Induction Time and Diagnosis Time Active Information Acquisition
Ashish Kapoor, Eric Horvitz
There has been a clear distinction between induction or training time and diagno sis time active information acquisition. While active learning during induction focuses on acquiring data that promises to provide the best classification model , the goal at diagnosis time focuses completely on next features to observe abou t the test case at hand in order to make better predictions about the case. We i ntroduce a model and inferential methods that breaks this distinction. The metho ds can be used to extend case libraries under a budget but, more fundamentally, provide a framework for guiding agents to collect data under scarce resources, f ocused by diagnostic challenges. This extension to active learning leads to a ne w class of policies for real-time diagnosis, where recommended information-gathe ring sequences include actions that simultaneously seek new data for the case at hand and for cases in the training set.
****

An Integer Projected Fixed Point Method for Graph Matching and MAP Inference
Marius Leordeanu, Martial Hebert, Rahul Sukthankar
Graph matching and MAP inference are essential problems in computer vision and m achine learning. We introduce a novel algorithm that can accommodate both proble

ms and solve them efficiently. Recent graph matching algorithms are based on a general quadratic programming formulation, that takes in consideration both unary and second-order terms reflecting the similarities in local appearance as well as in the pairwise geometric relationships between the matched features. In this case the problem is NP-hard and a lot of effort has been spent in finding efficiently approximate solutions by relaxing the constraints of the original problem. Most algorithms find optimal continuous solutions of the modified problem, ignoring during the optimization the original discrete constraints. The continuous solution is quickly binarized at the end, but very little attention is put into this final discretization step. In this paper we argue that the stage in which a discrete solution is found is crucial for good performance. We propose an efficient algorithm, with climbing and convergence properties, that optimizes in the discrete domain the quadratic score, and it gives excellent results either by itself or by starting from the solution returned by any graph matching algorithm. In practice it outperforms state-or-the art algorithms and it also significantly improves their performance if used in combination. When applied to MAP inference, the algorithm is a parallel extension of Iterated Conditional Modes (ICM) with climbing and convergence properties that make it a compelling alternative to the sequential ICM. In our experiments on MAP inference our algorithm proved its effectiveness by outperforming ICM and Max-Product Belief Propagation.

****

Efficient Bregman Range Search

Lawrence Cayton

We develop an algorithm for efficient range search when the notion of dissimilarity is given by a Bregman divergence. The range search task is to return all points in a potentially large database that are within some specified distance of a query. It arises in many learning algorithms such as locally-weighted regression, kernel density estimation, neighborhood graph-based algorithms, and in tasks like outlier detection and information retrieval. In metric spaces, efficient range search-like algorithms based on spatial data structures have been deployed on a variety of statistical tasks. Here we describe the first algorithm for range search for an arbitrary Bregman divergence. This broad class of dissimilarity measures includes the relative entropy, Mahalanobis distance, Itakura-Saito divergence, and a variety of matrix divergences. Metric methods cannot be directly applied since Bregman divergences do not in general satisfy the triangle inequality. We derive geometric properties of Bregman divergences that yield an efficient algorithm for range search based on a recently proposed space decomposition for Bregman divergences.

****

Complexity of Decentralized Control: Special Cases

Martin Allen, Shlomo Zilberstein

The worst-case complexity of general decentralized POMDPs, which are equivalent to partially observable stochastic games (POSGs) is very high, both for the cooperative and competitive cases. Some reductions in complexity have been achieved by exploiting independence relations in some models. We show that these results are somewhat limited: when these independence assumptions are relaxed in very small ways, complexity returns to that of the general case.

****