Emergence of Multiplication in a Biophysical Model of a Wide-Field Visual Neuron for Computing Object Approaches: Dynamics, Peaks, & Fits
Matthias Keil
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.
****
Action-Gap Phenomenon in Reinforcement Learning
Amir-massoud Farahmand
Many practitioners of reinforcement learning problems have observed that oftentimes the performance of the agent reaches very close to the optimal performance even though the estimated (action-)value function is still far from the optimal one. The goal of this paper is to explain and formalize this phenomenon by introducing the concept of the action-gap regularity. As a typical result, we prove that for an agent following the greedy policy (\hat{\pi}) with respect to an action-value function (\hat{Q}), the performance loss (E[V^(X) - V^{\hat{X}} (X)]) is upper bounded by (O(|| \hat{Q} - Q^||_\infty^{1+\zeta})), in which (\zeta >= 0) is the parameter quantifying the action-gap regularity. For (\zeta > 0), our results indicate smaller performance loss compared to what previous analyses had suggested. Finally, we show how this regularity affects the performance of the family of approximate value iteration algorithms.
****
Learning Higher-Order Graph Structure with Features by Structure Penalty
Shilin Ding, Grace Wahba, Jerry Zhu
In discrete undirected graphical models, the conditional independence of node labels Y is specified by the graph structure. We study the case where there is another input random vector X (e.g. observed features) such that the distribution P (Y | X) is determined by functions of X that characterize the (higher-order) interactions among the Y ’s. The main contribution of this paper is to learn the graph structure and the functions conditioned on X at the same time. We prove that discrete undirected graphical models with feature X are equivalent to mul- tivariate discrete models. The reparameterization of the potential functions in graphical models by conditional log odds ratios of the latter offers advantages in representation of the conditional independence structure. The functional spaces can be flexibly determined by kernels. Additionally, we impose a Structure Lasso (SLasso) penalty on groups of functions to learn the graph structure. These groups with overlaps are designed to enforce hierarchical function selection. In this way, we are able to shrink higher order interactions to obtain a sparse graph structure.
****
Efficient Methods for Overlapping Group Lasso
Lei Yuan, Jun Liu, Jieping Ye
The group Lasso is an extension of the Lasso for feature selection on (predefined) non-overlapping groups of features. The non-overlapping group structure limits its applicability in practice. There have been several recent attempts to study a more general formulation, where groups of features are given, potentially with overlaps between the groups. The resulting optimization is, however, much more challenging to solve due to the group overlaps. In this paper, we consider the efficient optimization of the overlapping group Lasso penalized problem. We reveal several key properties of the proximal operator associated with the overlapping group Lasso, and compute the proximal operator by solving the smooth and convex dual problem, which allows the use of the gradient descent type of algorithms for the optimization. We have performed empirical evaluations using both synthetic and the breast cancer gene expression data set, which consists of 8,141 genes organized into (overlapping) gene sets. Experimental results show that the proposed algorithm is more efficient than existing state-of-the-art algorithms.
****
Priors over Recurrent Continuous Time Processes
Ardavan Saeedi, Alexandre Bouchard-côté

We introduce the Gamma-Exponential Process (GEP), a prior over a large family of continuous time stochastic processes. A hierarchical version of this prior (HGEP; the Hierarchical GEP) yields a useful model for analyzing complex time series. Models based on HGEPs display many attractive properties: conjugacy, exchangeability and closed-form predictive distribution for the waiting times, and exact Gibbs updates for the time scale parameters. After establishing these properties, we show how posterior inference can be carried efficiently using Particle MCMC methods [1]. This yields a MCMC algorithm that can resample entire sequences atomically while avoiding the complications of introducing slice and stick auxiliary variables of the beam sampler [2]. We applied our model to the problem of estimating the disease progression in multiple sclerosis [3], and to RNA evolutionary modeling [4]. In both domains, we found that our model outperformed the standard rate matrix estimation approach.

****

The Kernel Beta Process

Lu Ren, Yingjian Wang, Lawrence Carin, David Dunson

A new Le ■vy process prior is proposed for an uncountable collection of covariate- dependent feature-learning measures; the model is called the kernel beta process (KBP). Available covariates are handled efficiently via the kernel construction, with covariates assumed observed with each data sample (“customer”), and latent covariates learned for each feature (“dish”). Each customer selects dishes from an infinite buffet, in a manner analogous to the beta process, with the added constraint that a customer first decides probabilistically whether to “consider” a dish, based on the distance in covariate space between the customer and dish. If a customer does consider a particular dish, that dish is then selected probabilistically as in the beta process. The beta process is recovered as a limiting case of the KBP. An efficient Gibbs sampler is developed for computations, and state-of-the-art results are presented for image processing and music analysis tasks.

****

Lower Bounds for Passive and Active Learning

Maxim Raginsky, Alexander Rakhlin

We develop unified information-theoretic machinery for deriving lower bounds for passive and active learning schemes. Our bounds involve the so-called Alexander's capacity function. The supremum of this function has been recently rediscovered by Hanneke in the context of active learning under the name of "disagreement coefficient." For passive learning, our lower bounds match the upper bounds of Gine and Koltchinskii up to constants and generalize analogous results of Massart and Nedelec. For active learning, we provide first known lower bounds based on the capacity function rather than the disagreement coefficient.

****

k-NN Regression Adapts to Local Intrinsic Dimension

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

****

Variational Learning for Recurrent Spiking Networks

Danilo Rezende, Daan Wierstra, Wulfram Gerstner

We derive a plausible learning rule updating the synaptic efficacies for feedforward, feedback and lateral connections between observed and latent neurons. Operating in the context of a generative model for distributions of spike sequences, the learning mechanism is derived from variational inference principles. The synaptic plasticity rules found are interesting in that they are strongly reminiscent of experimentally found results on Spike Time Dependent Plasticity, and in that they differ for excitatory and inhibitory neurons. A simulation confirms the method's applicability to learning both stationary and temporal spike patterns.

****

Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound
Iasonas Kokkinos
In this work we use Branch-and-Bound (BB) to efficiently detect objects with deformable part models. Instead of evaluating the classifier score exhaustively over image locations and scales, we use BB to focus on promising image locations. The core problem is to compute bounds that accommodate part deformations; for this we adapt the Dual Trees data structure to our problem. We evaluate our approach using Mixture-of-Deformable Part Models. We obtain exactly the same results but are 10-20 times faster on average. We also develop a multiple-object detection variation of the system, where hypotheses for 20 categories are inserted in a common priority queue. For the problem of finding the strongest category in an image this results in up to a 100-fold speedup.
****
Probabilistic Modeling of Dependencies Among Visual Short-Term Memory Representations
Emin Orhan, Robert Jacobs
Extensive evidence suggests that items are not encoded independently in visual short-term memory (VSTM). However, previous research has not quantitatively considered how the encoding of an item influences the encoding of other items. Here, we model the dependencies among VSTM representations using a multivariate Gaussian distribution with a stimulus-dependent mean and covariance matrix. We report the results of an experiment designed to determine the specific form of the stimulus-dependence of the mean and the covariance matrix. We find that the magnitude of the covariance between the representations of two items is a monotonically decreasing function of the difference between the items' feature values, similar to a Gaussian process with a distance-dependent, stationary kernel function. We further show that this type of covariance function can be explained as a natural consequence of encoding multiple stimuli in a population of neurons with correlated responses.
****
Bayesian Bias Mitigation for Crowdsourcing
Fabian L. Wauthier, Michael Jordan
Biased labelers are a systemic problem in crowdsourcing, and a comprehensive toolbox for handling their responses is still being developed. A typical crowdsourcing application can be divided into three steps: data collection, data curation, and learning. At present these steps are often treated separately. We present Bayesian Bias Mitigation for Crowdsourcing (BBMC), a Bayesian model to unify all three. Most data curation methods account for the {\it effects} of labeler bias by modeling all labels as coming from a single latent truth. Our model captures the {\it sources} of bias by describing labelers as influenced by shared random effects. This approach can account for more complex bias patterns that arise in ambiguous or hard labeling tasks and allows us to merge data curation and learning into a single computation. Active learning integrates data collection with learning, but is commonly considered infeasible with Gibbs sampling inference. We propose a general approximation strategy for Markov chains to efficiently quantify the effect of a perturbation on the stationary distribution and specialize this approach to active learning. Experiments show BBMC to outperform many common heuristics.
****
Phase transition in the family of p-resistances
Morteza Alamgir, Ulrike Luxburg
We study the family of p-resistances on graphs for $p \geq 1$. This family generalizes the standard resistance distance. We prove that for any fixed graph, for p=1, the p-resistance coincides with the shortest path distance, for p=2 it coincides with the standard resistance distance, and for $p \to \infty$ it converges to the inverse of the minimal s-t-cut in the graph. Secondly, we consider the special case of random geometric graphs (such as k-nearest neighbor graphs) when the number n of vertices in the graph tends to infinity. We prove that an interesting phase-transition takes place. There exist two critical thresholds p^* and p^* such that if p < p^, then the p-resistance depends on meaningful global properties of the

graph, whereas if p > p^, it only depends on trivial local quantities and does n ot convey any useful information. We can explicitly compute the critical values: p^* = 1 + 1/(d-1) and p^ = 1 + 1/(d-2) where d is the dimension of the underlyi ng space (we believe that the fact that there is a small gap between p^* and p^* is an artifact of our proofs. We also relate our findings to Laplacian regulari zation and suggest to use q-Laplacians as regularizers, where q satisfies 1/p^ + 1/q = 1.

****

On the Analysis of Multi-Channel Neural Spike Data

Bo Chen, David Carlson, Lawrence Carin

Nonparametric Bayesian methods are developed for analysis of multi-channel spik e-train data, with the feature learning and spike sorting performed jointly. Th e feature learning and sorting are performed simultaneously across all channels. Dictionary learning is implemented via the beta-Bernoulli process, with spike sorting performed via the dynamic hierarchical Dirichlet process (dHDP), with these two models coupled. The dHDP is augmented to eliminate refractoryperiod v iolations, it allows the “appearance” and “disappearance” of neurons over time, and it models smooth variation in the spike statistics.

****

Hashing Algorithms for Large-Scale Learning

Ping Li, Anshumali Shrivastava, Joshua Moore, Arnd König

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.

****

Active learning of neural response functions with Gaussian processes

Mijung Park, Greg Horwitz, Jonathan Pillow

A sizable literature has focused on the problem of estimating a low-dimensional feature space capturing a neuron's stimulus sensitivity. However, comparatively little work has addressed the problem of estimating the nonlinear function from feature space to a neuron's output spike rate. Here, we use a Gaussian process (GP) prior over the infinite-dimensional space of nonlinear functions to obtain Bayesian estimates of the "nonlinearity" in the linear-nonlinear-Poisson (LNP) e ncoding model. This offers flexibility, robustness, and computational tractabili ty compared to traditional methods (e.g., parametric forms, histograms, cubic sp lines). Most importantly, we develop a framework for optimal experimental desig n based on uncertainty sampling. This involves adaptively selecting stimuli to characterize the nonlinearity with as little experimental data as possible, and relies on a method for rapidly updating hyperparameters using the Laplace approx imation. We apply these methods to data from color-tuned neurons in macaque V1. We estimate nonlinearities in the 3D space of cone contrasts, which reveal tha t V1 combines cone inputs in a highly nonlinear manner. With simulated experimen ts, we show that optimal design substantially reduces the amount of data require d to estimate this nonlinear combination rule.

****

Nonstandard Interpretations of Probabilistic Programs for Efficient Inference

David Wingate, Noah Goodman, Andreas Stuhlmueller, Jeffrey Siskind

Probabilistic programming languages allow modelers to specify a stochastic pr ocess using syntax that resembles modern programming languages. Because the program is in machine-readable format, a variety of techniques from compiler design and program analysis can be used to examine the structure of the distr ibution represented by the probabilistic program. We show how nonstandard in terpretations of probabilistic programs can be used to craft efficient infere nce algorithms: information about the structure of a distribution (such as gr adients or dependencies) is generated as a monad-like side computation while executing the program. These interpretations can be easily coded using speci al-purpose objects and operator overloading. We implement two examples of no nstandard interpretations in two different languages, and use them as buildin g blocks to construct inference algorithms: automatic differentiation, which

enables gradient based methods, and provenance tracking, which enables efficient construction of global proposals.

****

The Impact of Unlabeled Patterns in Rademacher Complexity Theory for Kernel Classifiers

Luca Oneto, Davide Anguita, Alessandro Ghio, Sandro Ridella

We derive here new generalization bounds, based on Rademacher Complexity theory, for model selection and error estimation of linear (kernel) classifiers, which exploit the availability of unlabeled samples. In particular, two results are obtained: the first one shows that, using the unlabeled samples, the confidence term of the conventional bound can be reduced by a factor of three; the second one shows that the unlabeled samples can be used to obtain much tighter bounds, by building localized versions of the hypothesis class containing the optimal classifier.

****

Regularized Laplacian Estimation and Fast Eigenvector Approximation

Patrick Perry, Michael W. Mahoney

Recently, Mahoney and Orecchia demonstrated that popular diffusion-based procedures to compute a quick approximation to the first nontrivial eigenvector of a data graph Laplacian exactly solve certain regularized Semi-Definite Programs (SDPs). In this paper, we extend that result by providing a statistical interpretation of their approximation procedure. Our interpretation will be analogous to the manner in which l2-regularized or l1-regularized l2 regression (often called Ridge regression and Lasso regression, respectively) can be interpreted in terms of a Gaussian prior or a Laplace prior, respectively, on the coefficient vector of the regression problem. Our framework will imply that the solutions to the Mahoney-Orecchia regularized SDP can be interpreted as regularized estimates of the pseudoinverse of the graph Laplacian. Conversely, it will imply that the solution to this regularized estimation problem can be computed very quickly by running, e.g., the fast diffusion-based PageRank procedure for computing an approximation to the first nontrivial eigenvector of the graph Laplacian. Empirical results are also provided to illustrate the manner in which approximate eigenvector computation implicitly performs statistical regularization, relative to running the corresponding exact algorithm.

****

The Doubly Correlated Nonparametric Topic Model

Dae Kim, Erik Sudderth

Topic models are learned via a statistical model of variation within document collections, but designed to extract meaningful semantic structure. Desirable traits include the ability to incorporate annotations or metadata associated with documents; the discovery of correlated patterns of topic usage; and the avoidance of parametric assumptions, such as manual specification of the number of topics. We propose a doubly correlated nonparametric topic (DCNT) model, the first model to simultaneously capture all three of these properties. The DCNT models metadata via a flexible, Gaussian regression on arbitrary input features; correlations via a scalable square-root covariance representation; and nonparametric selection from an unbounded series of potential topics via a stick-breaking construction. We validate the semantic structure and predictive performance of the DCNT using a corpus of NIPS documents annotated by various metadata.

****

Generalized Lasso based Approximation of Sparse Coding for Visual Recognition

Nobuyuki Morioka, Shin'ichi Satoh

Sparse coding, a method of explaining sensory data with as few dictionary bases as possible, has attracted much attention in computer vision. For visual object category recognition, L1 regularized sparse coding is combined with spatial pyramid representation to obtain state-of-the-art performance. However, because of its iterative optimization, applying sparse coding onto every local feature descriptor extracted from an image database can become a major bottleneck. To overcome this computational challenge, this paper presents "Generalized Lasso based Approximation of Sparse coding" (GLAS). By representing the distribution of sparse

coefficients with slice transform, we fit a piece-wise linear mapping function w ith generalized lasso. We also propose an efficient post-refinement procedure to perform mutual inhibition between bases which is essential for an overcomplete setting. The experiments show that GLAS obtains comparable performance to L1 reg ularized sparse coding, yet achieves significant speed up demonstrating its effe ctiveness for large-scale visual recognition problems.
****

SpaRCS: Recovering low-rank and sparse matrices from compressive measurements
Andrew Waters, Aswin Sankaranarayanan, Richard Baraniuk
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Learning Patient-Specific Cancer Survival Distributions as a Sequence of Depende nt Regressors
Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, Vickie Baracos
An accurate model of patient survival time can help in the treatment and care of cancer patients. The common practice of providing survival time estimates based only on population averages for the site and stage of cancer ignores many impor tant individual differences among patients. In this paper, we propose a local re gression method for learning patient-specific survival time distribution based o n patient attributes such as blood tests and clinical assessments. When tested o n a cohort of more than 2000 cancer patients, our method gives survival time pre dictions that are much more accurate than popular survival analysis models such as the Cox and Aalen regression models. Our results also show that using patient -specific attributes can reduce the prediction error on survival time by as much as 20% when compared to using cancer site and stage only.
****

Prediction strategies without loss
Michael Kapralov, Rina Panigrahy
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.
****

Maximum Margin Multi-Instance Learning
Hua Wang, Heng Huang, Farhad Kamangar, Feiping Nie, Chris Ding
Multi-instance learning (MIL) considers input as bags of instances, in which lab els are assigned to the bags. MIL is useful in many real-world applications. For example, in image categorization semantic meanings (labels) of an image mostly arise from its regions (instances) instead of the entire image (bag). Existing M IL methods typically build their models using the Bag-to-Bag (B2B) distance, whi ch are often computationally expensive and may not truly reflect the semantic si milarities. To tackle this, in this paper we approach MIL problems from a new pe rspective using the Class-to-Bag (C2B) distance, which directly assesses the rel ationships between the classes and the bags. Taking into account the two major c hallenges in MIL, high heterogeneity on data and weak label association, we prop ose a novel Maximum Margin Multi-Instance Learning (M3 I) approach to parameteri ze the C2B distance by introducing the class specific distance metrics and the l ocally adaptive significance coefficients. We apply our new approach to the auto matic image categorization tasks on three (one single-label and two multilabel) benchmark data sets. Extensive experiments have demonstrated promising results t hat validate the proposed method.
****

Anatomically Constrained Decoding of Finger Flexion from Electrocorticographic S ignals
Zuoguan Wang, Gerwin Schalk, Qiang Ji
Brain-computer interfaces (BCIs) use brain signals to convey a user's intent. So me BCI approaches begin by decoding kinematic parameters of movements from brain

signals, and then proceed to using these signals, in absence of movements, to a llow a user to control an output. Recent results have shown that electrocorticog raphic (ECoG) recordings from the surface of the brain in humans can give inform ation about kinematic parameters (e.g., hand velocity or finger flexion). The de coding approaches in these demonstrations usually employed classical classificat ion/regression algorithms that derive a linear mapping between brain signals and outputs. However, they typically only incorporate little prior information abou t the target kinematic parameter. In this paper, we show that different types of anatomical constraints that govern finger flexion can be exploited in this cont ext. Specifically, we incorporate these constraints in the construction, structu re, and the probabilistic functions of a switched non-parametric dynamic system (SNDS) model. We then apply the resulting SNDS decoder to infer the flexion of i ndividual fingers from the same ECoG dataset used in a recent study. Our results show that the application of the proposed model, which incorporates anatomical constraints, improves decoding performance compared to the results in the previo us work. Thus, the results presented in this paper may ultimately lead to neural ly controlled hand prostheses with full fine-grained finger articulation.

****

Efficient coding of natural images with a population of noisy Linear-Nonlinear n eurons

Yan Karklin, Eero Simoncelli

Efficient coding provides a powerful principle for explaining early sensory codi ng. Most attempts to test this principle have been limited to linear, noiseless models, and when applied to natural images, have yielded oriented filters consis tent with responses in primary visual cortex. Here we show that an efficient cod ing model that incorporates biologically realistic ingredients input and output noise, nonlinear response functions, and a metabolic cost on the firing rate p redicts receptive fields and response nonlinearities similar to those observed i n the retina. Specifically, we develop numerical methods for simultaneously lear ning the linear filters and response nonlinearities of a population of model neu rons, so as to maximize information transmission subject to metabolic costs. Whe n applied to an ensemble of natural images, the method yields filters that are c enter-surround and nonlinearities that are rectifying. The filters are organized into two populations, with On- and Off-centers, which independently tile the vi sual space. As observed in the primate retina, the Off-center neurons are more n umerous and have filters with smaller spatial extent. In the absence of noise, o ur method reduces to a generalized version of independent components analysis, w ith an adapted nonlinear "contrast" function; in this case, the optimal filters are localized and oriented.

****

Multilinear Subspace Regression: An Orthogonal Tensor Decomposition Approach

Qibin Zhao, Cesar F. Caiafa, Danilo Mandic, Liqing Zhang, Tonio Ball, Andreas Sc hulze-bonhage, Andrzej Cichocki

A multilinear subspace regression model based on so called latent variable decom position is introduced. Unlike standard regression methods which typically emplo y matrix (2D) data representations followed by vector subspace transformations, the proposed approach uses tensor subspace transformations to model common laten t variables across both the independent and dependent data. The proposed approac h aims to maximize the correlation between the so derived latent variables and i s shown to be suitable for the prediction of multidimensional dependent data fro m multidimensional independent data, where for the estimation of the latent vari ables we introduce an algorithm based on Multilinear Singular Value Decompositio n (MSVD) on a specially defined cross-covariance tensor. It is next shown that i n this way we are also able to unify the existing Partial Least Squares (PLS) an d N-way PLS regression algorithms within the same framework. Simulations on benc hmark synthetic data confirm the advantages of the proposed approach, in terms o f its predictive ability and robustness, especially for small sample sizes. The potential of the proposed technique is further illustrated on a real world task of the decoding of human intracranial electrocorticogram (ECoG) from a simultane ously recorded scalp electroencephalograph (EEG).

****
Large-Scale Sparse Principal Component Analysis with Application to Text Data
Youwei Zhang, Laurent Ghaoui
Sparse PCA provides a linear combination of small number of features that maximizes variance across data. Although Sparse PCA has apparent advantages compared to PCA, such as better interpretability, it is generally thought to be computationally much more expensive. In this paper, we demonstrate the surprising fact that sparse PCA can be easier than PCA in practice, and that it can be reliably applied to very large data sets. This comes from a rigorous feature elimination pre-processing result, coupled with the favorable fact that features in real-life data typically have exponentially decreasing variances, which allows for many features to be eliminated. We introduce a fast block coordinate ascent algorithm with much better computational complexity than the existing first-order ones. We provide experimental results obtained on text corpora involving millions of documents and hundreds of thousands of features. These results illustrate how Sparse PCA can help organize a large corpus of text data in a user-interpretable way, providing an attractive alternative approach to topic models.
****
Nearest Neighbor based Greedy Coordinate Descent
Inderjit Dhillon, Pradeep Ravikumar, Ambuj Tewari
Increasingly, optimization problems in machine learning, especially those arising from high-dimensional statistical estimation, have a large number of variables. Modern statistical estimators developed over the past decade have statistical or sample complexity that depends only weakly on the number of parameters when there is some structure to the problem, such as sparsity. A central question is whether similar advances can be made in their computational complexity as well. In this paper, we propose strategies that indicate that such advances can indeed be made. In particular, we investigate the greedy coordinate descent algorithm, and note that performing the greedy step efficiently weakens the costly dependence on the problem size provided the solution is sparse. We then propose a suite of methods that perform these greedy steps efficiently by a reduction to nearest neighbor search. We also devise a more amenable form of greedy descent for composite non-smooth objectives; as well as several approximate variants of such greedy descent. We develop a practical implementation of our algorithm that combines greedy coordinate descent with locality sensitive hashing. Without tuning the latter data structure, we are not only able to significantly speed up the vanilla greedy method, but also outperform cyclic descent when the problem size becomes large. Our results indicate the effectiveness of our nearest neighbor strategies, and also point to many open questions regarding the development of computational geometric techniques tailored towards first-order optimization methods.
****
Convergent Bounds on the Euclidean Distance
Yoonho Hwang, Hee-kap Ahn
Given a set V of n vectors in d-dimensional space, we provide an efficient method for computing quality upper and lower bounds of the Euclidean distances between a pair of the vectors in V . For this purpose, we define a distance measure, called the MS-distance, by using the mean and the standard deviation values of vectors in V . Once we compute the mean and the standard deviation values of vectors in V in O(dn) time, the MS-distance between them provides upper and lower bounds of Euclidean distance between a pair of vectors in V in constant time. Furthermore, these bounds can be refined further such that they converge monotonically to the exact Euclidean distance within d refinement steps. We also provide an analysis on a random sequence of refinement steps which can justify why MS-distance should be refined to provide very tight bounds in a few steps of a typical sequence. The MS-distance can be used to various problems where the Euclidean distance is used to measure the proximity or similarity between objects. We provide experimental results on the nearest and the farthest neighbor searches.
****
Video Annotation and Tracking with Active Learning

Carl Vondrick, Deva Ramanan
We introduce a novel active learning framework for video annotation. By judicio
usly choosing which frames a user should annotate, we can obtain highly accurat
e tracks with minimal user effort. We cast this problem as one of active learn
ing, and show that we can obtain excellent performance by querying frames that,
 if annotated, would produce a large expected change in the estimated object tr
ack. We implement a constrained tracker and compute the expected change for put
ative annotations with efficient dynamic programming algorithms. We demonstrat
e our framework on four datasets, including two benchmark datasets constructed
with key frame annotations obtained by Amazon Mechanical Turk. Our results indi
cate that we could obtain equivalent labels for a small fraction of the origina
l cost.
****

PiCoDes: Learning a Compact Code for Novel-Category Recognition
Alessandro Bergamo, Lorenzo Torresani, Andrew Fitzgibbon
We introduce PiCoDes: a very compact image descriptor which nevertheless allows
high performance on object category recognition. In particular, we address novel
-category recognition: the task of defining indexing structures and image repres
entations which enable a large collection of images to be searched for an object
 category that was not known when the index was built. Instead, the training ima
ges defining the category are supplied at query time. We explicitly learn descri
ptors of a given length (from as small as 16 bytes per image) which have good ob
ject-recognition performance. In contrast to previous work in the domain of obje
ct recognition, we do not choose an arbitrary intermediate representation, but e
xplicitly learn short codes. In contrast to previous approaches to learn compact
 codes, we optimize explicitly for (an upper bound on) classification performanc
e. Optimization directly for binary features is difficult and nonconvex, but we
present an alternation scheme and convex upper bound which demonstrate excellent
 performance in practice. PiCoDes of 256 bytes match the accuracy of the current
 best known classifier for the Caltech256 benchmark, but they decrease the datab
ase storage size by a factor of 100 and speed-up the training and testing of nov
el classes by orders of magnitude.
****

Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Repre
sentation
Zhouchen Lin, Risheng Liu, Zhixun Su
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.
****

Sparse Filtering
Jiquan Ngiam, Zhenghao Chen, Sonia Bhaskar, Pang Koh, Andrew Ng
Unsupervised feature learning has been shown to be effective at learning represe
ntations that perform well on image, video and audio classification. However, ma
ny existing feature learning algorithms are hard to use and require extensive hy
perparameter tuning. In this work, we present sparse filtering, a simple new alg
orithm which is efficient and only has one hyperparameter, the number of feature
s to learn. In contrast to most other feature learning methods, sparse filterin
g does not explicitly attempt to construct a model of the data distribution. Ins
tead, it optimizes a simple cost function -- the sparsity of L2-normalized featu
res -- which can easily be implemented in a few lines of MATLAB code. Sparse fil
tering scales gracefully to handle high-dimensional inputs, and can also be used
 to learn meaningful features in additional layers with greedy layer-wise stacki
ng. We evaluate sparse filtering on natural images, object classification (STL-1
0), and phone classification (TIMIT), and show that our method works well on a r
ange of different modalities.
****

Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts
Matthias Hein, Simon Setzer

Spectral clustering is based on the spectral relaxation of the normalized/ratio graph cut criterion. While the spectral relaxation is known to be loose, it has been shown recently that a non-linear eigenproblem yields a tight relaxation of the Cheeger cut. In this paper, we extend this result considerably by providing a characterization of all balanced graph cuts which allow for a tight relaxation. Although the resulting optimization problems are non-convex and non-smooth, we provide an efficient first-order scheme which scales to large graphs. Moreover, our approach comes with the quality guarantee that given any partition as initialization the algorithm either outputs a better partition or it stops immediately.

****

Why The Brain Separates Face Recognition From Object Recognition

Joel Z. Leibo, Jim Mutch, Tomaso Poggio

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some of these regions are organized in a hierarchical manner with viewpoint-specific cells projecting to downstream viewpoint-invariant identity-specific cells (Freiwald and Tsao 2010). A separate computational line of reasoning leads to the claim that some transformations of visual inputs that preserve viewed object identity are class-specific. In particular, the 2D images evoked by a face undergoing a 3D rotation are not produced by the same image transformation (2D) that would produce the images evoked by an object of another class undergoing the same 3D rotation. However, within the class of faces, knowledge of the image transformation evoked by 3D rotation can be reliably transferred from previously viewed faces to help identify a novel face at a new viewpoint. We show, through computational simulations, that an architecture which applies this method of gaining invariance to class-specific transformations is effective when restricted to faces and fails spectacularly when applied across object classes. We argue here that in order to accomplish viewpoint-invariant face identification from a single example view, visual cortex must separate the circuitry involved in discounting 3D rotations of faces from the generic circuitry involved in processing other objects. The resulting model of the ventral stream of visual cortex is consistent with the recent physiology results showing the hierarchical organization of the face processing network.

****

Analytical Results for the Error in Filtering of Gaussian Processes

Alex K. Susemihl, Ron Meir, Manfred Opper

Bayesian filtering of stochastic stimuli has received a great deal of attention re- cently. It has been applied to describe the way in which biological systems dy- namically represent and make decisions about the environment. There have been no exact results for the error in the biologically plausible setting of inference on point process, however. We present an exact analysis of the evolution of the mean- squared error in a state estimation task using Gaussian-tuned point processes as sensors. This allows us to study the dynamics of the error of an optimal Bayesian decoder, providing insights into the limits obtainable in this task. This is done for Markovian and a class of non-Markovian Gaussian processes. We find that there is an optimal tuning width for which the error is minimized. This leads to a char- acterization of the optimal encoding for the setting as a function of the statistics of the stimulus, providing a mathematically sound primer for an ecological theory of sensory processing.

****

Active Learning with a Drifting Distribution

Liu Yang

We study the problem of active learning in a stream-based setting, allowing the distribution of the examples to change over time. We prove upper bounds on the number of prediction mistakes and number of label requests for established disagreement-based active learning algorithms, both in the realizable case and under Tsybakov noise. We further prove minimax lower bounds for this problem.

****

Evaluating the inverse decision-making approach to preference learning
Alan Jern, Christopher Lucas, Charles Kemp
Psychologists have recently begun to develop computational accounts of how peopl e infer others' preferences from their behavior. The inverse decision-making app roach proposes that people infer preferences by inverting a generative model of decision-making. Existing data sets, however, do not provide sufficient resoluti on to thoroughly evaluate this approach. We introduce a new preference learning task that provides a benchmark for evaluating computational accounts and use it to compare the inverse decision-making approach to a feature-based approach, whi ch relies on a discriminative combination of decision features. Our data support the inverse decision-making approach to preference learning. A basic principle of decision-making is that knowing people's preferences allows us to predict how they will behave: if you know your friend likes comedies and hates horror films , you can probably guess which of these options she will choose when she goes to the theater. Often, however, we do not know what other people like and we can o nly infer their preferences from their behavior. If you know that a different fr iend saw a comedy today, does that mean that he likes comedies in general? The c onclusion you draw will likely depend on what else was playing and what movie ch oices he has made in the past. A goal for social cognition research is to develo p a computational account of people's ability to infer others' preferences. One computational approach is based on inverse decision-making. This approach begins with a model of how someone's preferences lead to a decision. Then, this model is inverted to determine the most likely preferences that motivated an observed decision. An alternative approach might simply learn a functional mapping betwee n features of an observed decision and the preferences that motivated it. For in stance, in your friend's decision to see a comedy, perhaps the more movie option s he turned down, the more likely it is that he has a true preference for comedi es. The difference between the inverse decision-making approach and the feature-based approach maps onto the standard dichotomy between generative and discrimin ative models. Economists have developed an instance of the inverse decision-maki ng approach known as the multinomial logit model [1] that has been widely used t o infer consumer's preferences from their choices. This model has recently been explored as a psychological model [2, 3, 4], but there are few behavioral data s ets for evaluating it as a model of how people learn others' preferences. Additi onally, the data sets that do exist tend to be drawn from the developmental lite rature, which focuses on simple tasks that collect only one or two judgments fro m children [5, 6, 7]. The limitations of these data sets make it difficult to ev aluate the multinomial logit model with respect to alternative accounts of prefe rence learning like the feature-based approach. In this paper, we use data from a new experimental task that elicits a detailed set of preference judgments from a single participant in order to evaluate the predictions of several preference learning models from both the inverse decision-making and feature-based classes . Our task requires each participant to sort a large number of observed decision s on the basis of how strongly they indicate 1
****
Policy Gradient Coagent Networks
Philip S. Thomas
We present a novel class of actor-critic algorithms for actors consisting of set s of interacting modules. We present, analyze theoretically, and empirically eva luate an update rule for each module, which requires only local information: the module's input, output, and the TD error broadcast by a critic. Such updates ar e necessary when computation of compatible features becomes prohibitively diffic ult and are also desirable to increase the biological plausibility of reinforcem ent learning methods.
****
Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning
Joni Pajarinen, Jaakko Peltonen
Applications such as robot control and wireless communication require planning u nder uncertainty. Partially observable Markov decision processes (POMDPs) plan p olicies for single agents under uncertainty and their decentralized versions (DE

C-POMDPs) find a policy for multiple agents. The policy in infinite-horizon POMDP and DEC-POMDP problems has been represented as finite state controllers (FSCs). We introduce a novel class of periodic FSCs, composed of layers connected only to the previous and next layer. Our periodic FSC method finds a deterministic finite-horizon policy and converts it to an initial periodic infinite-horizon policy. This policy is optimized by a new infinite-horizon algorithm to yield deterministic periodic policies, and by a new expectation maximization algorithm to yield stochastic periodic policies. Our method yields better results than earlier planning methods and can compute larger solutions than with regular FSCs.
****

Spectral Methods for Learning Multivariate Latent Tree Structure
Animashree Anandkumar, Kamalika Chaudhuri, Daniel J. Hsu, Sham M. Kakade, Le Song, Tong Zhang
This work considers the problem of learning the structure of multivariate linear tree models, which include a variety of directed tree graphical models with continuous, discrete, and mixed latent variables such as linear-Gaussian models, hidden Markov models, Gaussian mixture models, and Markov evolutionary trees. The setting is one where we only have samples from certain observed variables in the tree, and our goal is to estimate the tree structure (i.e., the graph of how the underlying hidden variables are connected to each other and to the observed variables). We propose the Spectral Recursive Grouping algorithm, an efficient and simple bottom-up procedure for recovering the tree structure from independent samples of the observed variables. Our finite sample size bounds for exact recovery of the tree structure reveal certain natural dependencies on underlying statistical and structural properties of the underlying joint distribution. Furthermore, our sample complexity guarantees have no explicit dependence on the dimensionality of the observed variables, making the algorithm applicable to many high-dimensional settings. At the heart of our algorithm is a spectral quartet test for determining the relative topology of a quartet of variables from second-order statistics.
****

Query-Aware MCMC
Michael Wick, Andrew McCallum
Traditional approaches to probabilistic inference such as loopy belief propagation and Gibbs sampling typically compute marginals for it all the unobserved variables in a graphical model. However, in many real-world applications the user's interests are focused on a subset of the variables, specified by a query. In this case it would be wasteful to uniformly sample, say, one million variables when the query concerns only ten. In this paper we propose a query-specific approach to MCMC that accounts for the query variables and their generalized mutual information with neighboring variables in order to achieve higher computational efficiency. Surprisingly there has been almost no previous work on query-aware MCMC. We demonstrate the success of our approach with positive experimental results on a wide range of graphical models.
****

Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent
Benjamin Recht, Christopher Re, Stephen Wright, Feng Niu
Stochastic Gradient Descent (SGD) is a popular algorithm that can achieve state-of-the-art performance on a variety of machine learning tasks. Several researchers have recently proposed schemes to parallelize SGD, but all require performance-destroying memory locking and synchronization. This work aims to show using novel theoretical analysis, algorithms, and implementation that SGD can be implemented without any locking. We present an update scheme called Hogwild which allows processors access to shared memory with the possibility of overwriting each other's work. We show that when the associated optimization problem is sparse, meaning most gradient updates only modify small parts of the decision variable, then Hogwild achieves a nearly optimal rate of convergence. We demonstrate experimentally that Hogwild outperforms alternative schemes that use locking by an order of magnitude.
****

ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning
Quoc Le, Alexandre Karpenko, Jiquan Ngiam, Andrew Ng
Independent Components Analysis (ICA) and its variants have been successfully used for unsupervised feature learning. However, standard ICA requires an orthonormality constraint to be enforced, which makes it dif■cult to learn overcomplete features. In addition, ICA is sensitive to whitening. These properties make it challenging to scale ICA to high dimensional data. In this paper, we propose a robust soft reconstruction cost for ICA that allows us to learn highly overcomplete sparse features even on unwhitened data. Our formulation reveals formal connections between ICA and sparse autoencoders, which have previously been observed only empirically. Our algorithm can be used in conjunction with off-the-shelf fast unconstrained optimizers. We show that the soft reconstruction cost can also be used to prevent replicated features in tiled convolutional neural networks. Using our method to learn highly overcomplete sparse features and tiled convolutional neural networks, we obtain competitive performances on a wide variety of object recognition tasks. We achieve state-of-the-art test accuracies on the STL-10 and Hollywood2 datasets.
****
Sparse Recovery with Brownian Sensing
Alexandra Carpentier, Odalric-ambrym Maillard, Rémi Munos
We consider the problem of recovering the parameter alpha in R^K of a sparse function f, i.e. the number of non-zero entries of alpha is small compared to the number K of features, given noisy evaluations of f at a set of well-chosen sampling points. We introduce an additional randomisation process, called Brownian sensing, based on the computation of stochastic integrals, which produces a Gaussian sensing matrix, for which good recovery properties are proven independently on the number of sampling points N, even when the features are arbitrarily non-orthogonal. Under the assumption that f is Hölder continuous with exponent at least 1/2, we provide an estimate a of the parameter such that ||\alpha - a||2 = O(||eta||2\sqrt{N}), where eta is the observation noise. The method uses a set of sampling points uniformly distributed along a one-dimensional curve selected according to the features. We report numerical experiments illustrating our method.
****
Learning Anchor Planes for Classification
Ziming Zhang, Lubor Ladicky, Philip Torr, Amir Saffari
Local Coordinate Coding (LCC) [18] is a method for modeling functions of data lying on non-linear manifolds. It provides a set of anchor points which form a local coordinate system, such that each data point on the manifold can be approximated by a linear combination of its anchor points, and the linear weights become the local coordinate coding. In this paper we propose encoding data using orthogonal anchor planes, rather than anchor points. Our method needs only a few orthogonal anchor planes for coding, and it can linearize any (\alpha,\beta,p)-Lipschitz smooth nonlinear function with a fixed expected value of the upper-bound approximation error on any high dimensional data. In practice, the orthogonal coordinate system can be easily learned by minimizing this upper bound using singular value decomposition (SVD). We apply our method to model the coordinates locally in linear SVMs for classification tasks, and our experiment on MNIST shows that using only 50 anchor planes our method achieves 1.72% error rate, while LCC achieves 1.90% error rate using 4096 anchor points.
****
Ranking annotators for crowdsourced labeling tasks
Vikas C. Raykar, Shipeng Yu
With the advent of crowdsourcing services it has become quite cheap and reasonably effective to get a dataset labeled by multiple annotators in a short amount of time. Various methods have been proposed to estimate the consensus labels by correcting for the bias of annotators with different kinds of expertise. Often we have low quality annotators or spammers--annotators who assign labels randomly (e.g., without actually looking at the instance). Spammers can make the cost of acquiring labels very expensive and can potentially degrade the quality of the consensus labels. In this paper we formalize the notion of a spammer and define a

score which can be used to rank the annotators---with the spammers having a score close to zero and the good annotators having a high score close to one.
****
Metric Learning with Multiple Kernels
Jun Wang, Huyen T., Adam Woznica, Alexandros Kalousis
Metric learning has become a very active research field. The most popular representative--Mahalanobis metric learning--can be seen as learning a linear transformation and then computing the Euclidean metric in the transformed space. Since a linear transformation might not always be appropriate for a given learning problem, kernelized versions of various metric learning algorithms exist. However, the problem then becomes finding the appropriate kernel function. Multiple kernel learning addresses this limitation by learning a linear combination of a number of predefined kernels; this approach can be also readily used in the context of multiple-source learning to fuse different data sources. Surprisingly, and despite the extensive work on multiple kernel learning for SVMs, there has been no work in the area of metric learning with multiple kernel learning. In this paper we fill this gap and present a general approach for metric learning with multiple kernel learning. Our approach can be instantiated with different metric learning algorithms provided that they satisfy some constraints. Experimental evidence suggests that our approach outperforms metric learning with an unweighted kernel combination and metric learning with cross-validation based kernel selection.
****
A Brain-Machine Interface Operating with a Real-Time Spiking Neural Network Control Algorithm
Julie Dethier, Paul Nuyujukian, Chris Eliasmith, Terrence Stewart, Shauki Elasaad, Krishna V. Shenoy, Kwabena A. Boahen
Motor prostheses aim to restore function to disabled patients. Despite compelling proof of concept systems, barriers to clinical translation remain. One challenge is to develop a low-power, fully-implantable system that dissipates only minimal power so as not to damage tissue. To this end, we implemented a Kalman-filter based decoder via a spiking neural network (SNN) and tested it in brain-machine interface (BMI) experiments with a rhesus monkey. The Kalman filter was trained to predict the arm’s velocity and mapped on to the SNN using the Neural Engineer- ing Framework (NEF). A 2,000-neuron embedded Matlab SNN implementation runs in real-time and its closed-loop performance is quite comparable to that of the standard Kalman filter. The success of this closed-loop decoder holds promise for hardware SNN implementations of statistical signal processing algorithms on neuromorphic chips, which may offer power savings necessary to overcome a major obstacle to the successful clinical translation of neural motor prostheses.
****
Understanding the Intrinsic Memorability of Images
Phillip Isola, Devi Parikh, Antonio Torralba, Aude Oliva
Artists, advertisers, and photographers are routinely presented with the task of creating an image that a viewer will remember. While it may seem like image memorability is purely subjective, recent work shows that it is not an inexplicable phenomenon: variation in memorability of images is consistent across subjects, suggesting that some images are intrinsically more memorable than others, independent of a subjects' contexts and biases. In this paper, we used the publicly available memorability dataset of Isola et al., and augmented the object and scene annotations with interpretable spatial, content, and aesthetic image properties. We used a feature-selection scheme with desirable explaining-away properties to determine a compact set of attributes that characterizes the memorability of any individual image. We find that images of enclosed spaces containing people with visible faces are memorable, while images of vistas and peaceful scenes are not. Contrary to popular belief, unusual or aesthetically pleasing scenes do not tend to be highly memorable. This work represents one of the first attempts at understanding intrinsic image memorability, and opens a new domain of investigation at the interface between human cognition and computer vision.
****

Two is better than one: distinct roles for familiarity and recollection in retri
eving palimpsest memories
Cristina Savin, Peter Dayan, Máté Lengyel
Storing a new pattern in a palimpsest memory system comes at the cost of interfe
ring with the memory traces of previously stored items. Knowing the age of a pat
tern thus becomes critical for recalling it faithfully. This implies that there
should be a tight coupling between estimates of age, as a form of familiarity, a
nd the neural dynamics of recollection, something which current theories omit. U
sing a normative model of autoassociative memory, we show that a dual memory sys
tem, consisting of two interacting modules for familiarity and recollection, has
 best performance for both recollection and recognition. This finding provides a
 new window onto actively contentious psychological and neural aspects of recogn
ition memory.
****

Automated Refinement of Bayes Networks' Parameters based on Test Ordering Constr
aints
Omar Khan, Pascal Poupart, John-mark Agosta
In this paper, we derive a method to refine a Bayes network diagnostic model by
exploiting constraints implied by expert decisions on test ordering. At each ste
p, the expert executes an evidence gathering test, which suggests the test's rel
ative diagnostic value. We demonstrate that consistency with an expert's test se
lection leads to non-convex constraints on the model parameters. We incorporate
 these constraints by augmenting the network with nodes that represent the const
raint likelihoods. Gibbs sampling, stochastic hill climbing and greedy search al
gorithms are proposed to find a MAP estimate that takes into account test orderi
ng constraints and any data available. We demonstrate our approach on diagnostic
 sessions from a manufacturing scenario.
****

Structure Learning for Optimization
Shulin Yang, Ali Rahimi
We describe a family of global optimization procedures that automatically decomp
ose optimization problems into smaller loosely coupled problems, then combine th
e solutions of these with message passing algorithms. We show empirically that
these methods excel in avoiding local minima and produce better solutions with f
ewer function evaluations than existing global optimization methods. To develop
 these methods, we introduce a notion of coupling between variables of optimizat
ion that generalizes the notion of coupling that arises from factoring functions
 into terms that involve small subsets of the variables. It therefore subsumes t
he notion of independence between random variables in statistics, sparseness of
the Hessian in nonlinear optimization, and the generalized distributive law. Des
pite being more general, this notion of coupling is easier to verify empirically
 -- making structure estimation easy -- yet it allows us to migrate well-establi
shed inference methods on graphical models to the setting of global optimization
.
****

Multiclass Boosting: Theory and Algorithms
Mohammad Saberian, Nuno Vasconcelos
The problem of multiclass boosting is considered. A new framework,based on mult
i-dimensional codewords and predictors is introduced. The optimal set of codewor
ds is derived, and a margin enforcing loss proposed. The resulting risk is minim
ized by gradient descent on a multidimensional functional space. Two algorithms
are proposed: 1) CD-MCBoost, based on coordinate descent, updates one predictor
component at a time, 2) GD-MCBoost, based on gradient descent, updates all compo
nents jointly. The algorithms differ in the weak learners that they support but
are both shown to be 1) Bayes consistent, 2) margin enforcing, and 3) convergent
 to the global minimum of the risk. They also reduce to AdaBoost when there are
only two classes. Experiments show that both methods outperform previous multicl
ass boosting approaches on a number of datasets.
****

Composite Multiclass Losses

Elodie Vernet, Mark D. Reid, Robert C. Williamson
We consider loss functions for multiclass prediction problems. We show when a multiclass loss can be expressed as a proper composite loss'', which is the composition of a proper loss and a link function. We extend existing results for binary losses to multiclass losses. We determine the stationarity condition, Bregman representation, order-sensitivity, existence and uniqueness of the composite representation for multiclass losses. We also show that the integral representation for binary proper losses can not be extended to multiclass losses. We subsume existing results onclassification calibration'' by relating it to properness. We draw conclusions concerning the design of multiclass losses.
****

Scalable Training of Mixture Models via Coresets
Dan Feldman, Matthew Faulkner, Andreas Krause
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.
****

Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance
Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, Peter Gehler
We address the challenging task of decoupling material properties from lighting properties given a single image. In the last two decades virtually all works have concentrated on exploiting edge information to address this problem. We take a different route by introducing a new prior on reflectance, that models reflectance values as being drawn from a sparse set of basis colors. This results in a Random Field model with global, latent variables (basis colors) and pixel-accurate output reflectance values. We show that without edge information high-quality results can be achieved, that are on par with methods exploiting this source of information. Finally, we present competitive results by integrating an additional edge model. We believe that our approach is a solid starting point for future development in this domain.
****

Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation
Cho-jui Hsieh, Inderjit Dhillon, Pradeep Ravikumar, Mátyás Sustik
The L_1 regularized Gaussian maximum likelihood estimator has been shown to have strong statistical guarantees in recovering a sparse inverse covariance matrix, or alternatively the underlying graph structure of a Gaussian Markov Random Field, from very limited samples. We propose a novel algorithm for solving the resulting optimization problem which is a regularized log-determinant program. In contrast to other state-of-the-art methods that largely use first order gradient information, our algorithm is based on Newton's method and employs a quadratic approximation, but with some modifications that leverage the structure of the sparse Gaussian MLE problem. We show that our method is superlinearly convergent, and also present experimental results using synthetic and real application data that demonstrate the considerable improvements in performance of our method when compared to other state-of-the-art methods.
****

Testing a Bayesian Measure of Representativeness Using a Large Image Database
Joshua T. Abbott, Katherine A. Heller, Zoubin Ghahramani, Thomas Griffiths
How do people determine which elements of a set are most representative of that set? We extend an existing Bayesian measure of representativeness, which indicates the representativeness of a sample from a distribution, to define a measure of the representativeness of an item to a set. We show that this measure is formally related to a machine learning method known as Bayesian Sets. Building on this connection, we derive an analytic expression for the representativeness of objects described by a sparse vector of binary features. We then apply this measure to a large database of images, using it to determine which images are the most representative members of different sets. Comparing the resulting predictions to human judgments of representativeness provides a test of this measure with natu

ralistic stimuli, and illustrates how databases that are more commonly used in c
omputer vision and machine learning can be used to evaluate psychological theori
es.
****

Dynamical segmentation of single trials from population neural data
Biljana Petreska, Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen Ryu,
 Krishna V. Shenoy, Maneesh Sahani
Simultaneous recordings of many neurons embedded within a recurrently-connected
 cortical network may provide concurrent views into the dynamical processes of t
hat network, and thus its computational function. In principle, these dynamics
might be identified by purely unsupervised, statistical means. Here, we show th
at a Hidden Switching Linear Dynamical Systems (HSLDS) model---in which multiple
 linear dynamical laws approximate a nonlinear and potentially non-stationary dy
namical process---is able to distinguish different dynamical regimes within sing
le-trial motor cortical activity associated with the preparation and initiation
of hand movements. The regimes are identified without reference to behavioural
or experimental epochs, but nonetheless transitions between them correlate stron
gly with external events whose timing may vary from trial to trial. The HSLDS m
odel also performs better than recent comparable models in predicting the firing
 rate of an isolated neuron based on the firing rates of others, suggesting that
 it captures more of the "shared variance" of the data. Thus, the method is ab
le to trace the dynamical processes underlying the coordinated evolution of netw
ork activity in a way that appears to reflect its computational role.
****

Approximating Semidefinite Programs in Sublinear Time
Dan Garber, Elad Hazan
In recent years semidefinite optimization has become a tool of major importance
in various optimization and machine learning problems. In many of these problems
 the amount of data in practice is so large that there is a constant need for fa
ster algorithms. In this work we present the first sublinear time approximation
 algorithm for semidefinite programs which we believe may be useful for such pro
blems in which the size of data may cause even linear time algorithms to have pr
ohibitive running times in practice. We present the algorithm and its analysis a
longside with some theoretical lower bounds and an improved algorithm for the sp
ecial problem of supervised learning of a distance metric.
****

Active Classification based on Value of Classifier
Tianshi Gao, Daphne Koller
Modern classification tasks usually involve many class labels and can be informe
d by a broad range of features. Many of these tasks are tackled by constructing
a set of classifiers, which are then applied at test time and then pieced togeth
er in a fixed procedure determined in advance or at training time. We present an
 active classification process at the test time, where each classifier in a larg
e ensemble is viewed as a potential observation that might inform our classifica
tion process. Observations are then selected dynamically based on previous obser
vations, using a value-theoretic computation that balances an estimate of the ex
pected classification gain from each observation as well as its computational co
st. The expected classification gain is computed using a probabilistic model tha
t uses the outcome from previous observations. This active classification proces
s is applied at test time for each individual test instance, resulting in an eff
icient instance-specific decision path. We demonstrate the benefit of the active
 scheme on various real-world datasets, and show that it can achieve comparable
or even higher classification accuracy at a fraction of the computational costs
of traditional methods.
****

Efficient Learning of Generalized Linear and Single Index Models with Isotonic R
egression
Sham M. Kakade, Varun Kanade, Ohad Shamir, Adam Kalai
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.
****

Co-regularized Multi-view Spectral Clustering
Abhishek Kumar, Piyush Rai, Hal Daume
In many clustering problems, we have access to multiple views of the data each
of which could be individually used for clustering. Exploiting information from
multiple views, one can hope to find a clustering that is more accurate than th
e ones obtained using the individual views. Since the true clustering would ass
ign a point to the same cluster irrespective of the view, we can approach this
problem by looking for clusterings that are consistent across the views, i.e.,
corresponding data points in each view should have same cluster membership. We
propose a spectral clustering framework that achieves this goal by co-regulariz
ing the clustering hypotheses, and propose two co-regularization schemes to acc
omplish this. Experimental comparisons with a number of baselines on two synthe
tic and three real-world datasets establish the efficacy of our proposed approa
ches.
****

A concave regularization technique for sparse mixture models
Martin Larsson, Johan Ugander
Latent variable mixture models are a powerful tool for exploring the structure i
n large datasets. A common challenge for interpreting such models is a desire to
impose sparsity, the natural assumption that each data point only contains few
latent features. Since mixture distributions are constrained in their L1 norm, t
ypical sparsity techniques based on L1 regularization become toothless, and conc
ave regularization becomes necessary. Unfortunately concave regularization typic
ally results in EM algorithms that must perform problematic non-concave M-step m
aximizations. In this work, we introduce a technique for circumventing this diff
iculty, using the so-called Mountain Pass Theorem to provide easily verifiable c
onditions under which the M-step is well-behaved despite the lacking concavity.
We also develop a correspondence between logarithmic regularization and what we
term the pseudo-Dirichlet distribution, a generalization of the ordinary Dirichl
et distribution well-suited for inducing sparsity. We demonstrate our approach o
n a text corpus, inferring a sparse topic mixture model for 2,406 weblogs.
****

Image Parsing with Stochastic Scene Grammar
Yibiao Zhao, Song-chun Zhu
This paper proposes a parsing algorithm for scene understanding which includes f
our aspects: computing 3D scene layout, detecting 3D objects (e.g. furniture), d
etecting 2D faces (windows, doors etc.), and segmenting background. In contrast
to previous scene labeling work that applied discriminative classifiers to pixel
s (or super-pixels), we use a generative Stochastic Scene Grammar (SSG). This gr
ammar represents the compositional structures of visual entities from scene cate
gories, 3D foreground/background, 2D faces, to 1D lines. The grammar includes th
ree types of production rules and two types of contextual relations. Production
rules: (i) AND rules represent the decomposition of an entity into sub-parts; (i
i) OR rules represent the switching among sub-types of an entity; (iii) SET rule
s rep- resent an ensemble of visual entities. Contextual relations: (i) Cooperat
ive “+” relations represent positive links between binding entities, such as hin
ged faces of a object or aligned boxes; (ii) Competitive “-” relations represent
s negative links between competing entities, such as mutually exclusive boxes. W
e design an efficient MCMC inference algorithm, namely Hierarchical cluster samp
ling, to search in the large solution space of scene configurations. The algorit
hm has two stages: (i) Clustering: It forms all possible higher-level structures
(clusters) from lower-level entities by production rules and contextual relatio
ns. (ii) Sampling: It jumps between alternative structures (clusters) in each la
yer of the hierarchy to find the most probable configuration (represented by a p
arse tree). In our experiment, we demonstrate the superiority of our algorithm o
ver existing methods on public dataset. In addition, our approach achieves riche
r structures in the parse tree.

****
Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection
Richard Socher, Eric Huang, Jeffrey Pennin, Christopher D. Manning, Andrew Ng
Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two statements is needed. We introduce a method for paraphrase detection based on recursive autoencoders (RAE). Our unsupervised RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences. Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size. We introduce a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation is then used as input to a classifier. Our method outperforms other state-of-the-art approaches on the challenging MSRP paraphrase corpus.
****
Trace Lasso: a trace norm regularization for correlated designs
Edouard Grave, Guillaume R. Obozinski, 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-authors prior to requesting a name change in the electronic proceedings.
****
Linear Submodular Bandits and their Application to Diversified Retrieval
Yisong Yue, Carlos Guestrin
Diversified retrieval and online learning are two core research areas in the design of modern information retrieval systems.In this paper, we propose the linear submodular bandits problem, which is an online learning setting for optimizing a general class of feature-rich submodular utility models for diversified retrieval. We present an algorithm, called LSBGREEDY, and prove that it efficiently converges to a near-optimal model. As a case study, we applied our approach to the setting of personalized news recommendation, where the system must recommend small sets of news articles selected from tens of thousands of available articles each day. In a live user study, we found that LSBGREEDY significantly outperforms existing online learning approaches.
****
Learning Eigenvectors for Free
Wouter M. Koolen, Wojciech Kotlowski, Manfred K. K. Warmuth
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.
****
Neural Reconstruction with Approximate Message Passing (NeuRAMP)
Alyson K. Fletcher, Sundeep Rangan, Lav R. Varshney, Aniruddha Bhargava
Many functional descriptions of spiking neurons assume a cascade structure where inputs are passed through an initial linear filtering stage that produces a low-dimensional signal that drives subsequent nonlinear stages. This paper presents a novel and systematic parameter estimation procedure for such models and applies the method to two neural estimation problems: (i) compressed-sensing based neural mapping from multi-neuron excitation, and (ii) estimation of neural receptive yields in sensory neurons. The proposed estimation algorithm models the neurons via a graphical model and then estimates the parameters in the model using a recently-developed generalized approximate message passing (GAMP) method. The GAMP method is based on Gaussian approximations of loopy belief propagation. In the neural connectivity problem, the GAMP-based method is shown to be computational efficient, provides a more exact modeling of the sparsity, can incorporate nonlinearities in the output and significantly outperforms previous compressed-sensing methods. For the receptive field estimation, the GAMP method can also exploit inherent structured sparsity in the linear weights. The method is vali

dated on estimation of linear nonlinear Poisson (LNP) cascade models for recepti
ve fields of salamander retinal ganglion cells.
****

Bayesian Partitioning of Large-Scale Distance Data
David Adametz, Volker Roth
A Bayesian approach to partitioning distance matrices is presented. It is inspir
ed by the 'Translation-Invariant Wishart-Dirichlet' process (TIWD) in (Vogt et a
l., 2010) and shares a number of advantageous properties like the fully probabil
istic nature of the inference model, automatic selection of the number of cluste
rs and applicability in semi-supervised settings. In addition, our method (which
 we call 'fastTIWD') overcomes the main shortcoming of the original TIWD, namely
 its high computational costs. The fastTIWD reduces the workload in each iterati
on of a Gibbs sampler from O(n^3) in the TIWD to O(n^2). Our experiments show th
at this cost reduction does not compromise the quality of the inferred partition
s. With this new method it is now possible to 'mine' large relational datasets w
ith a probabilistic model, thereby automatically detecting new and potentially i
nteresting clusters.
****

Dimensionality Reduction Using the Sparse Linear Model
Ioannis Gkioulekas, Todd Zickler
We propose an approach for linear unsupervised dimensionality reduction, based o
n the sparse linear model that has been used to probabilistically interpret spar
se coding. We formulate an optimization problem for learning a linear projection
 from the original signal domain to a lower-dimensional one in a way that approx
imately preserves, in expectation, pairwise inner products in the sparse domain.
 We derive solutions to the problem, present nonlinear extensions, and discuss r
elations to compressed sensing. Our experiments using facial images, texture pat
ches, and images of object categories suggest that the approach can improve our
ability to recover meaningful structure in many classes of signals.
****

RTRMC: A Riemannian trust-region method for low-rank matrix completion
Nicolas Boumal, Pierre-antoine Absil
We consider large matrices of low rank. We address the problem of recovering suc
h matrices when most of the entries are unknown. Matrix completion finds applica
tions in recommender systems. In this setting, the rows of the matrix may corres
pond to items and the columns may correspond to users. The known entries are the
 ratings given by users to some items. The aim is to predict the unobserved rati
ngs. This problem is commonly stated in a constrained optimization framework. We
 follow an approach that exploits the geometry of the low-rank constraint to rec
ast the problem as an unconstrained optimization problem on the Grassmann manifo
ld. We then apply first- and second-order Riemannian trust-region methods to sol
ve it. The cost of each iteration is linear in the number of known entries. Our
methods, RTRMC 1 and 2, outperform state-of-the-art algorithms on a wide range o
f problem instances.
****

Complexity of Inference in Latent Dirichlet Allocation
David Sontag, Dan Roy
We consider the computational complexity of probabilistic inference in Latent Di
richlet Allocation (LDA). First, we study the problem of finding the maximum a
posteriori (MAP) assignment of topics to words, where the document's topic distr
ibution is integrated out. We show that, when the effective number of topics pe
r document is small, exact inference takes polynomial time. In contrast, we show
 that, when a document has a large number of topics, finding the MAP assignment
of topics to words in LDA is NP-hard. Next, we consider the problem of finding
the MAP topic distribution for a document, where the topic-word assignments are
integrated out. We show that this problem is also NP-hard. Finally, we briefly
 discuss the problem of sampling from the posterior, showing that this is NP-har
d in one restricted setting, but leaving open the general question.
****

A Denoising View of Matrix Completion

Weiran Wang, Miguel Carreira-Perpiñán, Zhengdong Lu
In matrix completion, we are given a matrix where the values of only some of the entries are present, and we want to reconstruct the missing ones. Much work has focused on the assumption that the data matrix has low rank. We propose a more general assumption based on denoising, so that we expect that the value of a missing entry can be predicted from the values of neighboring points. We propose a nonparametric version of denoising based on local, iterated averaging with mean-shift, possibly constrained to preserve local low-rank manifold structure. The few user parameters required (the denoising scale, number of neighbors and local dimensionality) and the number of iterations can be estimated by cross-validating the reconstruction error. Using our algorithms as a postprocessing step on an initial reconstruction (provided by e.g. a low-rank method), we show consistent improvements with synthetic, image and motion-capture data.
****

Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss
Joseph Keshet, David McAllester
We consider latent structural versions of probit loss and ramp loss. We show that these surrogate loss functions are consistent in the strong sense that for any feature map (finite or infinite dimensional) they yield predictors approaching the infimum task loss achievable by any linear predictor over the given features. We also give finite sample generalization bounds (convergence rates) for these loss functions. These bounds suggest that probit loss converges more rapidly. However, ramp loss is more easily optimized and may ultimately be more practical.
****

Budgeted Optimization with Concurrent Stochastic-Duration Experiments
Javad Azimi, Alan Fern, Xiaoli Fern
Budgeted optimization involves optimizing an unknown function that is costly to evaluate by requesting a limited number of function evaluations at intelligently selected inputs. Typical problem formulations assume that experiments are selected one at a time with a limited total number of experiments, which fail to capture important aspects of many real-world problems. This paper defines a novel problem formulation with the following important extensions: 1) allowing for concurrent experiments; 2) allowing for stochastic experiment durations; and 3) placing constraints on both the total number of experiments and the total experimental time. We develop both offline and online algorithms for selecting concurrent experiments in this new setting and provide experimental results on a number of optimization benchmarks. The results show that our algorithms produce highly effective schedules compared to natural baselines.
****

Data Skeletonization via Reeb Graphs
Xiaoyin Ge, Issam Safa, Mikhail Belkin, Yusu Wang
Recovering hidden structure from complex and noisy non-linear data is one of the most fundamental problems in machine learning and statistical inference. While such data is often high-dimensional, it is of interest to approximate it with a low-dimensional or even one-dimensional space, since many important aspects of data are often intrinsically low-dimensional. Furthermore, there are many scenarios where the underlying structure is graph-like, e.g, river/road networks or various trajectories. In this paper, we develop a framework to extract, as well as to simplify, a one-dimensional "skeleton" from unorganized data using the Reeb graph. Our algorithm is very simple, does not require complex optimizations and can be easily applied to unorganized high-dimensional data such as point clouds or proximity graphs. It can also represent arbitrary graph structures in the data. We also give theoretical results to justify our method. We provide a number of experiments to demonstrate the effectiveness and generality of our algorithm, including comparisons to existing methods, such as principal curves. We believe that the simplicity and practicality of our algorithm will help to promote skeleton graphs as a data analysis tool for a broad range of applications.
****

MAP Inference for Bayesian Inverse Reinforcement Learning
Jaedeug Choi, Kee-eung Kim
The difficulty in inverse reinforcement learning (IRL) arises in choosing the best reward function since there are typically an infinite number of reward functions that yield the given behaviour data as optimal. Using a Bayesian framework, we address this challenge by using the maximum a posteriori (MAP) estimation for the reward function, and show that most of the previous IRL algorithms can be modeled into our framework. We also present a gradient method for the MAP estimation based on the (sub)differentiability of the posterior distribution. We show the effectiveness of our approach by comparing the performance of the proposed method to those of the previous algorithms.
****
Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries
Zhen Xiang, Hao Xu, Peter J. Ramadge
Learning sparse representations on data adaptive dictionaries is a state-of-the-art method for modeling data. But when the dictionary is large and the data dimension is high, it is a computationally challenging problem. We explore three aspects of the problem. First, we derive new, greatly improved screening tests that quickly identify codewords that are guaranteed to have zero weights. Second, we study the properties of random projections in the context of learning sparse representations. Finally, we develop a hierarchical framework that uses incremental random projections and screening to learn, in small stages, a hierarchically structured dictionary for sparse representations. Empirical results show that our framework can learn informative hierarchical sparse representations more efficiently.
****
Efficient Online Learning via Randomized Rounding
Nicolò Cesa-bianchi, Ohad Shamir
Most online algorithms used in machine learning today are based on variants of mirror descent or follow-the-leader. In this paper, we present an online algorithm based on a completely different approach, which combines ``random playout'' and randomized rounding of loss subgradients. As an application of our approach, we provide the first computationally efficient online algorithm for collaborative filtering with trace-norm constrained matrices. As a second application, we solve an open question linking batch learning and transductive online learning.
****
Spatial distance dependent Chinese restaurant processes for image segmentation
Soumya Ghosh, Andrei Ungureanu, Erik Sudderth, David Blei
The distance dependent Chinese restaurant process (ddCRP) was recently introduced to accommodate random partitions of non-exchangeable data. The ddCRP clusters data in a biased way: each data point is more likely to be clustered with other data that are near it in an external sense. This paper examines the ddCRP in a spatial setting with the goal of natural image segmentation. We explore the biases of the spatial ddCRP model and propose a novel hierarchical extension better suited for producing "human-like" segmentations. We then study the sensitivity of the models to various distance and appearance hyperparameters, and provide the first rigorous comparison of nonparametric Bayesian models in the image segmentation domain. On unsupervised image segmentation, we demonstrate that similar performance to existing nonparametric Bayesian models is possible with substantially simpler models and algorithms.
****
History distribution matching method for predicting effectiveness of HIV combination therapies
Jasmina Bogojeska
This paper presents an approach that predicts the effectiveness of HIV combination therapies by simultaneously addressing several problems affecting the available HIV clinical data sets: the different treatment backgrounds of the samples, the uneven representation of the levels of therapy experience, the missing treatment history information, the uneven therapy representation and the unbalanced th

erapy outcome representation. The computational validation on clinical data shows that, compared to the most commonly used approach that does not account for the issues mentioned above, our model has significantly higher predictive power. This is especially true for samples stemming from patients with longer treatment history and samples associated with rare therapies. Furthermore, our approach is at least as powerful for the remaining samples.
****

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

Eric Moulines, Francis Bach

We consider the minimization of a convex objective function defined on a Hilbert space, which is only available through unbiased estimates of its gradients. This problem includes standard machine learning algorithms such as kernel logistic regression and least-squares regression, and is commonly referred to as a stochastic approximation problem in the operations research community. We provide a non-asymptotic analysis of the convergence of two well-known algorithms, stochastic gradient descent (a.k.a.~Robbins-Monro algorithm) as well as a simple modification where iterates are averaged (a.k.a.~Polyak-Ruppert averaging). Our analysis suggests that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the lack of strong convexity or the setting of the proportionality constant. This situation is remedied when using slower decays together with averaging, robustly leading to the optimal rate of convergence. We illustrate our theoretical results with simulations on synthetic and standard datasets.
****

A Non-Parametric Approach to Dynamic Programming

Oliver Kroemer, Jan Peters

In this paper, we consider the problem of policy evaluation for continuous-state systems. We present a non-parametric approach to policy evaluation, which uses kernel density estimation to represent the system. The true form of the value function for this model can be determined, and can be computed using Galerkin's method. Furthermore, we also present a unified view of several well-known policy evaluation methods. In particular, we show that the same Galerkin method can be used to derive Least-Squares Temporal Difference learning, Kernelized Temporal Difference learning, and a discrete-state Dynamic Programming solution, as well as our proposed method. In a numerical evaluation of these algorithms, the proposed approach performed better than the other methods.
****

Extracting Speaker-Specific Information with a Regularized Siamese Deep Network

Ke Chen, Ahmad Salman

Speech conveys different yet mixed information ranging from linguistic to speaker-specific components, and each of them should be exclusively used in a specific task. However, it is extremely difficult to extract a specific information component given the fact that nearly all existing acoustic representations carry all types of speech information. Thus, the use of the same representation in both speech and speaker recognition hinders a system from producing better performance due to interference of irrelevant information. In this paper, we present a deep neural architecture to extract speaker-specific information from MFCCs. As a result, a multi-objective loss function is proposed for learning speaker-specific characteristics and regularization via normalizing interference of non-speaker related information and avoiding information loss. With LDC benchmark corpora and a Chinese speech corpus, we demonstrate that a resultant speaker-specific representation is insensitive to text/languages spoken and environmental mismatches and hence outperforms MFCCs and other state-of-the-art techniques in speaker recognition. We discuss relevant issues and relate our approach to previous work.
****

Variance Penalizing AdaBoost

Pannagadatta Shivaswamy, Tony Jebara

This paper proposes a novel boosting algorithm called VadaBoost which is motivat

ed by recent empirical Bernstein bounds. VadaBoost iteratively minimizes a cost function that balances the sample mean and the sample variance of the exponential loss. Each step of the proposed algorithm minimizes the cost efficiently by providing weighted data to a weak learner rather than requiring a brute force evaluation of all possible weak learners. Thus, the proposed algorithm solves a key limitation of previous empirical Bernstein boosting methods which required brute force enumeration of all possible weak learners. Experimental results confirm that the new algorithm achieves the performance improvements of EBBoost yet goes beyond decision stumps to handle any weak learner. Significant performance gains are obtained over AdaBoost for arbitrary weak learners including decision trees (CART).

****

Autonomous Learning of Action Models for Planning

Neville Mehta, Prasad Tadepalli, Alan Fern

This paper introduces two new frameworks for learning action models for planning. In the mistake-bounded planning framework, the learner has access to a planner for the given model representation, a simulator, and a planning problem generator, and aims to learn a model with at most a polynomial number of faulty plans. In the planned exploration framework, the learner does not have access to a problem generator and must instead design its own problems, plan for them, and converge with at most a polynomial number of planning attempts. The paper reduces learning in these frameworks to concept learning with one-sided error and provides algorithms for successful learning in both frameworks. A specific family of hypothesis spaces is shown to be efficiently learnable in both the frameworks.

****

Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte, Frank Wood, Noemie Elhadad, Nicholas Bartlett

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

****

Agnostic Selective Classification

Yair Wiener, Ran El-Yaniv

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.

****

Additive Gaussian Processes

David K. Duvenaud, Hannes Nickisch, Carl Rasmussen

We introduce a Gaussian process model of functions which are additive. An additive function is one which decomposes into a sum of low-dimensional functions, each depending on only a subset of the input variables. Additive GPs generalize both Generalized Additive Models, and the standard GP models which use squared-exponential kernels. Hyperparameter learning in this model can be seen as Bayesian Hierarchical Kernel Learning (HKL). We introduce an expressive but tractable parameterization of the kernel function, which allows efficient evaluation of all input interaction terms, whose number is exponential in the input dimension. The additional structure discoverable by this model results in increased interpretability, as well as state-of-the-art predictive power in regression tasks.

****

A Collaborative Mechanism for Crowdsourcing Prediction Problems

Jacob D. Abernethy, Rafael Frongillo

Machine Learning competitions such as the Netflix Prize have proven reasonably s uccessful as a method of “crowdsourcing” prediction tasks. But these compe- titi ons have a number of weaknesses, particularly in the incentive structure they cr eate for the participants. We propose a new approach, called a Crowdsourced Lear ning Mechanism, in which participants collaboratively “learn” a hypothesis for a given prediction task. The approach draws heavily from the concept of a predict ion market, where traders bet on the likelihood of a future event. In our framew ork, the mechanism continues to publish the current hypothesis, and par- ticipan ts can modify this hypothesis by wagering on an update. The critical in- centive property is that a participant will profit an amount that scales according to h ow much her update improves performance on a released test set.

****

Sparse Bayesian Multi-Task Learning

Shengbo Guo, Onno Zoeter, Cédric Archambeau

We propose a new sparse Bayesian model for multi-task regression and classificat ion. The model is able to capture correlations between tasks, or more specifical ly a low-rank approximation of the covariance matrix, while being sparse in the features. We introduce a general family of group sparsity inducing priors based on matrix-variate Gaussian scale mixtures. We show the amount of sparsity can be learnt from the data by combining an approximate inference approach with type I I maximum likelihood estimation of the hyperparameters. Empirical evaluations on data sets from biology and vision demonstrate the applicability of the model, w here on both regression and classification tasks it achieves competitive predict ive performance compared to previously proposed methods.

****

Orthogonal Matching Pursuit with Replacement

Prateek Jain, Ambuj Tewari, Inderjit Dhillon

In this paper, we consider the problem of compressed sensing where the goal is t o recover almost all the sparse vectors using a small number of fixed linear mea surements. For this problem, we propose a novel partial hard-thresholding operat or leading to a general family of iterative algorithms. While one extreme of the family yields well known hard thresholding algorithms like ITI and HTP, the oth er end of the spectrum leads to a novel algorithm that we call Orthogonal Match ing Pursuit with Replacement (OMPR). OMPR, like the classic greedy algorithm OMP , adds exactly one coordinate to the support at each iteration, based on the cor relation with the current residual. However, unlike OMP, OMPR also removes one c oordinate from the support. This simple change allows us to prove the best known guarantees for OMPR in terms of the Restricted Isometry Property (a condition o n the measurement matrix). In contrast, OMP is known to have very weak performan ce guarantees under RIP. We also extend OMPR using locality sensitive hashing t o get OMPR-Hash, the first provably sub-linear (in dimensionality) algorithm for sparse recovery. Our proof techniques are novel and flexible enough to also pe rmit the tightest known analysis of popular iterative algorithms such as CoSaMP and Subspace Pursuit. We provide experimental results on large problems providi ng recovery for vectors of size up to million dimensions. We demonstrate that for large-scale problems our proposed methods are more robust and faster than th e existing methods.

****

High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessa ry Conditions

Animashree Anandkumar, Vincent Tan, Alan Willsky

We consider the problem of Ising and Gaussian graphical model selection given n i.i.d. samples from the model. We propose an efficient threshold-based algorithm for structure estimation based known as conditional mutual information test. This simple local algorithm requires only low-order statistics of the data a nd decides whether two nodes are neighbors in the unknown graph. Under som e transparent assumptions, we establish that the proposed algorithm is structura lly consistent (or sparsistent) when the number of samples scales as n= Omega(J {min}^{-4} log p), where p is the number of nodes and J{min} is the minimum edge potential. We also prove novel non-asymptotic necessary conditions for graphic

al model selection.
****
Optimal learning rates for least squares SVMs using Gaussian kernels
Mona Eberts, Ingo Steinwart
We prove a new oracle inequality for support vector machines with Gaussian RBF k ernels solving the regularized least squares regression problem. To this end, we apply the modulus of smoothness. With the help of the new oracle inequality we then derive learning rates that can also be achieved by a simple data-dependent parameter selection method. Finally, it turns out that our learning rates are as ymptotically optimal for regression functions satisfying certain standard smooth ness conditions.
****
Fast and Accurate k-means For Large Datasets
Michael Shindler, Alex Wong, Adam Meyerson
Clustering is a popular problem with many applications. We consider the k-means problem in the situation where the data is too large to be stored in main memory and must be accessed sequentially, such as from a disk, and where we must use a s little memory as possible. Our algorithm is based on recent theoretical result s, with significant improvements to make it practical. Our approach greatly simp li(cid:173) fies a recently developed algorithm, both in design and in analysis, and eliminates large constant factors in the approximation guarantee, the memor y requirements, and the running time. We then incorporate approximate nearest ne ighbor search to compute k-means in o(nk) (where n is the number of data points; note that com(cid:173) puting the cost, given a solution, takes 8(nk) time). We show that our algorithm compares favorably to existing algorithms - both theore tically and experimentally, thus providing state-of-the-art performance in both theory and practice.
****
Message-Passing for Approximate MAP Inference with Latent Variables
Jiarong Jiang, Piyush Rai, Hal Daume
We consider a general inference setting for discrete probabilistic graphical mod els where we seek maximum a posteriori (MAP) estimates for a subset of the rando m variables (max nodes), marginalizing over the rest (sum nodes). We present a h ybrid message-passing algorithm to accomplish this. The hybrid algorithm passes a mix of sum and max messages depending on the type of source node (sum or max) . We derive our algorithm by showing that it falls out as the solution of a part icular relaxation of a variational framework. We further show that the Expectat ion Maximization algorithm can be seen as an approximation to our algorithm. Exp erimental results on synthetic and real-world datasets, against several baseline s, demonstrate the efficacy of our proposed algorithm.
****
A More Powerful Two-Sample Test in High Dimensions using Random Projection
Miles Lopes, Laurent Jacob, Martin J. Wainwright
We consider the hypothesis testing problem of detecting a shift between the mean s of two multivariate normal distributions in the high-dimensional setting, allo wing for the data dimension p to exceed the sample size n. Our contribution is a new test statistic for the two-sample test of means that integrates a random pr ojection with the classical Hotelling T squared statistic. Working within a high - dimensional framework that allows (p,n) to tend to infinity, we first derive a n asymptotic power function for our test, and then provide sufficient conditions for it to achieve greater power than other state-of-the-art tests. Using ROC c urves generated from simulated data, we demonstrate superior performance against competing tests in the parameter regimes anticipated by our theoretical results . Lastly, we illustrate an advantage of our procedure with comparisons on a high -dimensional gene expression dataset involving the discrimination of different t ypes of cancer.
****
Kernel Bayes' Rule
Kenji Fukumizu, Le Song, Arthur Gretton
A nonparametric kernel-based method for realizing Bayes' rule is proposed, based

on kernel representations of probabilities in reproducing kernel Hilbert spaces . The prior and conditional probabilities are expressed as empirical kernel mean and covariance operators, respectively, and the kernel mean of the posterior di stribution is computed in the form of a weighted sample. The kernel Bayes' rule can be applied to a wide variety of Bayesian inference problems: we demonstrate Bayesian computation without likelihood, and filtering with a nonparametric sta te-space model. A consistency rate for the posterior estimate is established.

****

ShareBoost: Efficient multiclass learning with feature sharing

Shai Shalev-shwartz, Yonatan Wexler, Amnon Shashua

Multiclass prediction is the problem of classifying an object into a relevant target class. We consider the problem of learning a multiclass predictor th at uses only few features, and in particular, the number of used features sho uld increase sub-linearly with the number of possible classes. This implies t hat features should be shared by several classes. We describe and analyze the ShareBoost algorithm for learning a multiclass predictor that uses few share d features. We prove that ShareBoost efficiently finds a predictor that us es few shared features (if such a predictor exists) and that it has a small g eneralization error. We also describe how to use ShareBoost for learning a no n-linear predictor that has a fast evaluation time. In a series of experiment s with natural data sets we demonstrate the benefits of ShareBoost and evalua te its success relatively to other state-of-the-art approaches.

****

Heavy-tailed Distances for Gradient Based Image Descriptors

Yangqing Jia, Trevor Darrell

Many applications in computer vision measure the similarity between images or im age patches based on some statistics such as oriented gradients. These are often modeled implicitly or explicitly with a Gaussian noise assumption, leading to t he use of the Euclidean distance when comparing image descriptors. In this paper , we show that the statistics of gradient based image descriptors often follow a heavy-tailed distribution, which undermines any principled motivation for the u se of Euclidean distances. We advocate for the use of a distance measure based o n the likelihood ratio test with appropriate probabilistic models that fit the e mpirical data distribution. We instantiate this similarity measure with the Gamm a-compound-Laplace distribution, and show significant improvement over existing distance measures in the application of SIFT feature matching, at relatively low computational cost.

****

An Application of Tree-Structured Expectation Propagation for Channel Decoding

Pablo Olmos, Luis Salamanca, Juan Fuentes, Fernando Pérez-Cruz

We show an application of a tree structure for approximate inference in graphica l models using the expectation propagation algorithm. These approximations are t ypically used over graphs with short-range cycles. We demonstrate that these app roximations also help in sparse graphs with long-range loops, as the ones used i n coding theory to approach channel capacity. For asymptotically large sparse gr aph, the expectation propagation algorithm together with the tree structure yiel ds a completely disconnected approximation to the graphical model but, for for f inite-length practical sparse graphs, the tree structure approximation to the co de graph provides accurate estimates for the marginal of each variable.

****

PAC-Bayesian Analysis of Contextual Bandits

Yevgeny Seldin, Peter Auer, John Shawe-taylor, Ronald Ortner, François Laviolett e

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.

****

Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognit ion

Jia Deng, Sanjeev Satheesh, Alexander Berg, Fei Li
We present a novel approach to efficiently learn a label tree for large scale cl
assification with many classes. The key contribution of the approach is a techni
que to simultaneously determine the structure of the tree and learn the classifi
ers for each node in the tree. This approach also allows fine grained control ov
er the efficiency vs accuracy trade-off in designing a label tree, leading to mo
re balanced trees. Experiments are performed on large scale image classification
 with 10184 classes and 9 million images. We demonstrate significant improvement
s in test accuracy and efficiency with less training time and more balanced tree
s compared to the previous state of the art by Bengio et al.
****

Divide-and-Conquer Matrix Factorization
Lester Mackey, Michael Jordan, Ameet Talwalkar
This work introduces Divide-Factor-Combine (DFC), a parallel divide-and-conquer
framework for noisy matrix factorization. DFC divides a large-scale matrix fact
orization task into smaller subproblems, solves each subproblem in parallel usin
g an arbitrary base matrix factorization algorithm, and combines the subproblem
solutions using techniques from randomized matrix approximation. Our experiment
s with collaborative filtering, video background modeling, and simulated data de
monstrate the near-linear to super-linear speed-ups attainable with this approac
h. Moreover, our analysis shows that DFC enjoys high-probability recovery guara
ntees comparable to those of its base algorithm.
****

Non-conjugate Variational Message Passing for Multinomial and Binary Regression
David Knowles, Tom Minka
Variational Message Passing (VMP) is an algorithmic implementation of the Variat
ional Bayes (VB) method which applies only in the special case of conjugate expo
nential family models. We propose an extension to VMP, which we refer to as Non-
conjugate Variational Message Passing (NCVMP) which aims to alleviate this restr
iction while maintaining modularity, allowing choice in how expectations are cal
culated, and integrating into an existing message-passing framework: Infer.NET.
We demonstrate NCVMP on logistic binary and multinomial regression. In the multi
nomial case we introduce a novel variational bound for the softmax factor which
is tighter than other commonly used bounds whilst maintaining computational trac
tability.
****

Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara Berg
We develop and demonstrate automatic image description methods using a large cap
tioned photo collection. One contribution is our technique for the automatic co
llection of this new dataset -- performing a huge number of Flickr queries and t
hen filtering the noisy results down to 1 million images with associated visuall
y relevant captions. Such a collection allows us to approach the extremely chal
lenging problem of description generation using relatively simple non-parametric
 methods and produces surprisingly effective results. We also develop methods in
corporating many state of the art, but fairly noisy, estimates of image content
to produce even more pleasing results. Finally we introduce a new objective perf
ormance measure for image captioning.
****

Modelling Genetic Variations using Fragmentation-Coagulation Processes
Yee Teh, Charles Blundell, Lloyd Elliott
We propose a novel class of Bayesian nonparametric models for sequential data ca
lled fragmentation-coagulation processes (FCPs). FCPs model a set of sequences
using a partition-valued Markov process which evolves by splitting and merging
clusters. An FCP is exchangeable, projective, stationary and reversible, and it
s equilibrium distributions are given by the Chinese restaurant process. As opp
osed to hidden Markov models, FCPs allow for flexible modelling of the number of
 clusters, and they avoid label switching non-identifiability problems. We devel
op an efficient Gibbs sampler for FCPs which uses uniformization and the forward
-backward algorithm. Our development of FCPs is motivated by applications in po

pulation genetics, and we demonstrate the utility of FCPs on problems of genotyp
e imputation with phased and unphased SNP data.
****

Uniqueness of Belief Propagation on Signed Graphs
Yusuke Watanabe
While loopy Belief Propagation (LBP) has been utilized in a wide variety of appl
ications with empirical success, it comes with few theoretical guarantees. Espec
ially, if the interactions of random variables in a graphical model are strong,
the behaviors of the algorithm can be difficult to analyze due to underlying pha
se transitions. In this paper, we develop a novel approach to the uniqueness pro
blem of the LBP fixed point; our new “necessary and sufficient” condition is sta
ted in terms of graphs and signs, where the sign denotes the types (attractive/r
epulsive) of the interaction (i.e., compatibility function) on the edge. In all
previous works, uniqueness is guaranteed only in the situations where the streng
th of the interactions are “sufficiently” small in certain senses. In contrast,
our condition covers arbitrary strong interactions on the specified class of sig
ned graphs. The result of this paper is based on the recent theoretical advance
in the LBP algorithm; the connection with the graph zeta function.
****

Improving Topic Coherence with Regularized Topic Models
David Newman, Edwin V. Bonilla, Wray Buntine
Topic models have the potential to improve search and browsing by extracting use
ful semantic themes from web pages and other text documents. When learned topics
 are coherent and interpretable, they can be valuable for faceted browsing, resu
lts set diversity analysis, and document retrieval. However, when dealing with s
mall collections or noisy text (e.g. web search result snippets or blog posts),
learned topics can be less coherent, less interpretable, and less useful. To ove
rcome this, we propose two methods to regularize the learning of topic models. O
ur regularizers work by creating a structured prior over words that reflect broa
d patterns in the external data. Using thirteen datasets we show that both regul
arizers improve topic coherence and interpretability while learning a faithful r
epresentation of the collection of interest. Overall, this work makes topic mode
ls more useful across a broader range of text data.
****

Beating SGD: Learning SVMs in Sublinear Time
Elad Hazan, Tomer Koren, Nati Srebro
We present an optimization approach for linear SVMs based on a stochastic primal
-dual approach, where the primal step is akin to an importance-weighted SGD, and
 the dual step is a stochastic update on the importance weights. This yields an
 optimization method with a sublinear dependence on the training set size, and t
he first method for learning linear SVMs with runtime less then the size of the
training set required for learning!
****

Inferring spike-timing-dependent plasticity from spike train data
Ian Stevenson, Konrad Koerding
Synaptic plasticity underlies learning and is thus central for development, memo
ry, and recovery from injury. However, it is often difficult to detect changes i
n synaptic strength in vivo, since intracellular recordings are experimentally c
hallenging. Here we present two methods aimed at inferring changes in the coupli
ng between pairs of neurons from extracellularly recorded spike trains. First, u
sing a generalized bilinear model with Poisson output we estimate time-varying c
oupling assuming that all changes are spike-timing-dependent. This approach allo
ws model-based estimation of STDP modification functions from pairs of spike tra
ins. Then, using recursive point-process adaptive filtering methods we estimate
more general variation in coupling strength over time. Using simulations of neur
ons undergoing spike-timing dependent modification, we show that the true modifi
cation function can be recovered. Using multi-electrode data from motor cortex w
e then illustrate the use of this technique on in vivo data.
****

Bayesian Spike-Triggered Covariance Analysis

Il Memming Park, Jonathan Pillow
Neurons typically respond to a restricted number of stimulus features within the high-dimensional space of natural stimuli. Here we describe an explicit model-based interpretation of traditional estimators for a neuron's multi-dimensional feature space, which allows for several important generalizations and extensions. First, we show that traditional estimators based on the spike-triggered average (STA) and spike-triggered covariance (STC) can be formalized in terms of the "expected log-likelihood" of a Linear-Nonlinear-Poisson (LNP) model with Gaussian stimuli. This model-based formulation allows us to define maximum-likelihood and Bayesian estimators that are statistically consistent and efficient in a wider variety of settings, such as with naturalistic (non-Gaussian) stimuli. It also allows us to employ Bayesian methods for regularization, smoothing, sparsification, and model comparison, and provides Bayesian confidence intervals on model parameters. We describe an empirical Bayes method for selecting the number of features, and extend the model to accommodate an arbitrary elliptical nonlinear response function, which results in a more powerful and more flexible model for feature space inference. We validate these methods using neural data recorded extracellularly from macaque primary visual cortex.
****

Adaptive Hedge
Tim Erven, Wouter M. Koolen, Steven Rooij, Peter Grünwald
Most methods for decision-theoretic online learning are based on the Hedge algorithm, which takes a parameter called the learning rate. In most previous analyses the learning rate was carefully tuned to obtain optimal worst-case performance, leading to suboptimal performance on easy instances, for example when there exists an action that is significantly better than all others. We propose a new way of setting the learning rate, which adapts to the difficulty of the learning problem: in the worst case our procedure still guarantees optimal performance, but on easy instances it achieves much smaller regret. In particular, our adaptive method achieves constant regret in a probabilistic setting, when there exists an action that on average obtains strictly smaller loss than all other actions. We also provide a simulation study comparing our approach to existing methods.
****

Matrix Completion for Multi-label Image Classification
Ricardo Cabral, Fernando Torre, Joao P. Costeira, Alexandre Bernardino
Recently, image categorization has been an active research topic due to the urgent need to retrieve and browse digital images via semantic keywords. This paper formulates image categorization as a multi-label classification problem using recent advances in matrix completion. Under this setting, classification of testing data is posed as a problem of completing unknown label entries on a data matrix that concatenates training and testing features with training labels. We propose two convex algorithms for matrix completion based on a Rank Minimization criterion specifically tailored to visual data, and prove its convergence properties. A major advantage of our approach w.r.t. standard discriminative classification methods for image categorization is its robustness to outliers, background noise and partial occlusions both in the feature and label space. Experimental validation on several datasets shows how our method outperforms state-of-the-art algorithms, while effectively capturing semantic concepts of classes.
****

Continuous-Time Regression Models for Longitudinal Networks
Duy Vu, David Hunter, Padhraic Smyth, Arthur Asuncion
The development of statistical models for continuous-time longitudinal network data is of increasing interest in machine learning and social science. Leveraging ideas from survival and event history analysis, we introduce a continuous-time regression modeling framework for network event data that can incorporate both time-dependent network statistics and time-varying regression coefficients. We also develop an efficient inference scheme that allows our approach to scale to large networks. On synthetic and real-world data, empirical results demonstrate that the proposed inference approach can accurately estimate the coefficients of the regression model, which is useful for interpreting the evolution of the netwo

rk; furthermore, the learned model has systematically better predictive performa
nce compared to standard baseline methods.
****

Stochastic convex optimization with bandit feedback
Alekh Agarwal, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, Alexander Rakhlin
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.
****

Online Learning: Stochastic, Constrained, and Smoothed Adversaries
Alexander Rakhlin, Karthik Sridharan, Ambuj Tewari
Learning theory has largely focused on two main learning scenarios: the classica
l statistical setting where instances are drawn i.i.d. from a fixed distribution
, and the adversarial scenario whereby at every time step the worst instance is
revealed to the player. It can be argued that in the real world neither of these
 assumptions is reasonable. We define the minimax value of a game where the adve
rsary is restricted in his moves, capturing stochastic and non-stochastic assump
tions on data. Building on the sequential symmetrization approach, we define a n
otion of distribution-dependent Rademacher complexity for the spectrum of proble
ms ranging from i.i.d. to worst-case. The bounds let us immediately deduce varia
tion-type bounds. We study a smoothed online learning scenario and show that exp
onentially small amount of noise can make function classes with infinite Littles
tone dimension learnable.
****

Similarity-based Learning via Data Driven Embeddings
Purushottam Kar, Prateek Jain
We consider the problem of classification using similarity/distance functions ov
er data. Specifically, we propose a framework for defining the goodness of a (di
s)similarity function with respect to a given learning task and propose algorith
ms that have guaranteed generalization properties when working with such good fu
nctions. Our framework unifies and generalizes the frameworks proposed by (Balca
n-Blum 2006) and (Wang et al 2007). An attractive feature of our framework is it
s adaptability to data - we do not promote a fixed notion of goodness but rather
 let data dictate it. We show, by giving theoretical guarantees that the goodnes
s criterion best suited to a problem can itself be learned which makes our appro
ach applicable to a variety of domains and problems. We propose a landmarking-ba
sed approach to obtaining a classifier from such learned goodness criteria. We t
hen provide a novel diversity based heuristic to perform task-driven selection o
f landmark points instead of random selection. We demonstrate the effectiveness
of our goodness criteria learning method as well as the landmark selection heuri
stic on a variety of similarity-based learning datasets and benchmark UCI datase
ts on which our method consistently outperforms existing approaches by a signifi
cant margin.
****

Maximum Margin Multi-Label Structured Prediction
Christoph H. Lampert
We study multi-label prediction for structured output spaces, a problem that occ
urs, for example, in object detection in images, secondary structure prediction
in computational biology, and graph matching with symmetries. Conventional multi
-label classification techniques are typically not applicable in this situation,
 because they require explicit enumeration of the label space, which is infeasib
le in case of structured outputs. Relying on techniques originally designed for
single- label structured prediction, in particular structured support vector mac
hines, results in reduced prediction accuracy, or leads to infeasible optimizati
on problems. In this work we derive a maximum-margin training formulation for m
ulti-label structured prediction that remains computationally tractable while ac
hieving high prediction accuracy. It also shares most beneficial properties with
 single-label maximum-margin approaches, in particular a formulation as a convex
 optimization problem, efficient working set training, and PAC-Bayesian generali

zation bounds.
****
Active Ranking using Pairwise Comparisons
Kevin G. Jamieson, Robert Nowak
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.
****
Selecting Receptive Fields in Deep Networks
Adam Coates, Andrew Ng
Recent deep learning and unsupervised feature learning systems that learn from unlabeled data have achieved high performance in benchmarks by using extremely large architectures with many features (hidden units) at each layer. Unfortunately, for such large architectures the number of parameters usually grows quadratically in the width of the network, thus necessitating hand-coded "local receptive fields" that limit the number of connections from lower level features to higher ones (e.g., based on spatial locality). In this paper we propose a fast method to choose these connections that may be incorporated into a wide variety of unsupervised training methods. Specifically, we choose local receptive fields that group together those low-level features that are most similar to each other according to a pairwise similarity metric. This approach allows us to harness the advantages of local receptive fields (such as improved scalability, and reduced data requirements) when we do not know how to specify such receptive fields by hand or where our unsupervised training algorithm has no obvious generalization to a topographic setting. We produce results showing how this method allows us to use even simple unsupervised training algorithms to train successful multi-layered etworks that achieve state-of-the-art results on CIFAR and STL datasets: 82.0% and 60.1% accuracy, respectively.
****
Learning Auto-regressive Models from Sequence and Non-sequence Data
Tzu-kuo Huang, Jeff Schneider
Vector Auto-regressive models (VAR) are useful tools for analyzing time series data. In quite a few modern time series modelling tasks, the collection of reliable time series turns out to be a major challenge, either due to the slow progression of the dynamic process of interest, or inaccessibility of repetitive measurements of the same dynamic process over time. In those situations, however, we observe that it is often easier to collect a large amount of non-sequence samples, or snapshots of the dynamic process of interest. In this work, we assume a small amount of time series data are available, and propose methods to incorporate non-sequence data into penalized least-square estimation of VAR models. We consider non-sequence data as samples drawn from the stationary distribution of the underlying VAR model, and devise a novel penalization scheme based on the discrete-time Lyapunov equation concerning the covariance of the stationary distribution. Experiments on synthetic and video data demonstrate the effectiveness of the proposed methods.
****
Multi-View Learning of Word Embeddings via CCA
Paramveer Dhillon, Dean P. Foster, Lyle Ungar
Recently, there has been substantial interest in using large amounts of unlabeled data to learn word representations which can then be used as features in supervised classifiers for NLP tasks. However, most current approaches are slow to train, do not model context of the word, and lack theoretical grounding. In this paper, we present a new learning method, Low Rank Multi-View Learning (LR-MVL) which uses a fast spectral method to estimate low dimensional context-specific word representations from unlabeled data. These representation features can then be used with any supervised learner. LR-MVL is extremely fast, gives guaranteed convergence to a global optimum, is theoretically elegant, and achieves state-of-the-art performance on named entity recognition (NER) and chunking problems.
****

Projection onto A Nonnegative Max-Heap
Jun Liu, Liang Sun, 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.
****

Learning to Learn with Compound HD Models
Antonio Torralba, Joshua Tenenbaum, Russ R. Salakhutdinov
We introduce HD (or ``Hierarchical-Deep'') models, a new compositional learning architecture that integrates deep learning models with structured hierarchical Bayesian models. Specifically we show how we can learn a hierarchical Dirichlet process (HDP) prior over the activities of the top-level features in a Deep Boltzmann Machine (DBM). This compound HDP-DBM model learns to learn novel concepts from very few training examples, by learning low-level generic features, high-level features that capture correlations among low-level features, and a category hierarchy for sharing priors over the high-level features that are typical of different kinds of concepts. We present efficient learning and inference algorithms for the HDP-DBM model and show that it is able to learn new concepts from very few examples on CIFAR-100 object recognition, handwritten character recognition, and human motion capture datasets.
****

Object Detection with Grammar Models
Ross Girshick, Pedro Felzenszwalb, David McAllester
Compositional models provide an elegant formalism for representing the visual appearance of highly variable objects. While such models are appealing from a theoretical point of view, it has been difficult to demonstrate that they lead to performance advantages on challenging datasets. Here we develop a grammar model for person detection and show that it outperforms previous high-performance systems on the PASCAL benchmark. Our model represents people using a hierarchy of deformable parts, variable structure and an explicit model of occlusion for partially visible objects. To train the model, we introduce a new discriminative framework for learning structured prediction models from weakly-labeled data.
****

Inductive reasoning about chimeric creatures
Charles Kemp
Given one feature of a novel animal, humans readily make inferences about other features of the animal. For example, winged creatures often fly, and creatures that eat fish often live in the water. We explore the knowledge that supports these inferences and compare two approaches. The first approach proposes that humans rely on abstract representations of dependency relationships between features, and is formalized here as a graphical model. The second approach proposes that humans rely on specific knowledge of previously encountered animals, and is formalized here as a family of exemplar models. We evaluate these models using a task where participants reason about chimeras, or animals with pairs of features that have not previously been observed to co-occur. The results support the hypothesis that humans rely on explicit representations of relationships between features.
****

Empirical models of spiking in neural populations
Jakob H. Macke, Lars Buesing, John P. Cunningham, Byron M. Yu, Krishna V. Shenoy, Maneesh Sahani
Neurons in the neocortex code and compute as part of a locally interconnected population. Large-scale multi-electrode recording makes it possible to access these population processes empirically by fitting statistical models to unaveraged data. What statistical structure best describes the concurrent spiking of cells within a local network? We argue that in the cortex, where firing exhibits extensive correlations in both time and space and where a typical sample of neurons still reflects only a very small fraction of the local population, the most appro

priate model captures shared variability by a low-dimensional latent process evo lving with smooth dynamics, rather than by putative direct coupling. We test thi s claim by comparing a latent dynamical model with realistic spiking observatio ns to coupled generalised linear spike-response models (GLMs) using cortical re cordings. We find that the latent dynamical approach outperforms the GLM in term s of goodness-of-fit, and reproduces the temporal correlations in the data more accurately. We also compare models whose observations models are either derived from a Gaussian or point-process models, finding that the non-Gaussian model pro vides slightly better goodness-of-fit and more realistic population spike count s.

****

An Exact Algorithm for F-Measure Maximization

Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, Eyke Hüllermeier

The F-measure, originally introduced in information retrieval, is nowadays routi nely used as a performance metric for problems such as binary classification, mu lti-label classification, and structured output prediction. Optimizing this meas ure remains a statistically and computationally challenging problem, since no cl osed-form maximizer exists. Current algorithms are approximate and typically rel y on additional assumptions regarding the statistical distribution of the binary response variables. In this paper, we present an algorithm which is not only co mputationally efficient but also exact, regardless of the underlying distributio n. The algorithm requires only a quadratic number of parameters of the joint dis tribution (with respect to the number of binary responses). We illustrate its p ractical performance by means of experimental results for multi-label classifica tion.

****

Exploiting spatial overlap to efficiently compute appearance distances between i mage windows

Bogdan Alexe, Viviana Petrescu, Vittorio Ferrari

We present a computationally efficient technique to compute the distance of high -dimensional appearance descriptor vectors between image windows. The method ex ploits the relation between appearance distance and spatial overlap. We derive an upper bound on appearance distance given the spatial overlap of two windows i n an image, and use it to bound the distances of many pairs between two images. We propose algorithms that build on these basic operations to efficiently solv e tasks relevant to many computer vision applications, such as finding all pairs of windows between two images with distance smaller than a threshold, or findi ng the single pair with the smallest distance. In experiments on the PASCAL VOC 07 dataset, our algorithms accurately solve these problems while greatly reducin g the number of appearance distances computed, and achieve larger speedups than approximate nearest neighbour algorithms based on trees [18]and on hashing [21]. For example, our algorithm finds the most similar pair of windows between tw o images while computing only 1% of all distances on average.

****

Signal Estimation Under Random Time-Warpings and Nonlinear Signal Alignment

Sebastian Kurtek, Anuj Srivastava, Wei Wu

While signal estimation under random amplitudes, phase shifts, and additive nois e is studied frequently, the problem of estimating a deterministic signal under random time-warpings has been relatively unexplored. We present a novel framewor k for estimating the unknown signal that utilizes the action of the warping grou p to form an equivalence relation between signals. First, we derive an estimator for the equivalence class of the unknown signal using the notion of Karcher mea n on the quotient space of equivalence classes. This step requires the use of Fi sher-Rao Riemannian metric and a square-root representation of signals to enabl e computations of distances and means under this metric. Then, we define a notio n of the center of a class and show that the center of the estimated class is a consistent estimator of the underlying unknown signal. This estimation algorithm has many applications: (1)registration/alignment of functional data, (2) separa tion of phase/amplitude components of functional data, (3) joint demodulation an d carrier estimation, and (4) sparse modeling of functional data. Here we demons

trate only (1) and (2): Given signals are temporally aligned using nonlinear wa rpings and, thus, separated into their phase and amplitude components. The propo sed method for signal alignment is shown to have state of the art performance us ing Berkeley growth, handwritten signatures, and neuroscience spike train data.
****
Multi-armed bandits on implicit metric spaces
Aleksandrs Slivkins
The multi-armed bandit (MAB) setting is a useful abstraction of many online lear ning tasks which focuses on the trade-off between exploration and exploitation. In this setting, an online algorithm has a fixed set of alternatives ("arms"), a nd in each round it selects one arm and then observes the corresponding reward. While the case of small number of arms is by now well-understood, a lot of recen t work has focused on multi-armed bandits with (infinitely) many arms, where one needs to assume extra structure in order to make the problem tractable. In part icular, in the Lipschitz MAB problem there is an underlying similarity metric sp ace, known to the algorithm, such that any two arms that are close in this metri c space have similar payoffs. In this paper we consider the more realistic scen ario in which the metric space is implicit -- it is defined by the available str ucture but not revealed to the algorithm directly. Specifically, we assume that an algorithm is given a tree-based classification of arms. For any given problem instance such a classification implicitly defines a similarity metric space, bu t the numerical similarity information is not available to the algorithm. We pro vide an algorithm for this setting, whose performance guarantees (almost) match the best known guarantees for the corresponding instance of the Lipschitz MAB pr oblem.
****
Predicting response time and error rates in visual search
Bo Chen, Vidhya Navalpakkam, Pietro Perona
A model of human visual search is proposed. It predicts both response time (RT) and error rates (RT) as a function of image parameters such as target contrast a nd clutter. The model is an ideal observer, in that it optimizes the Bayes ratio of tar- get present vs target absent. The ratio is computed on the firing patte rn of V1/V2 neurons, modeled by Poisson distributions. The optimal mechanism for integrat- ing information over time is shown to be a ‘soft max’ of diffusions, computed over the visual field by ‘hypercolumns’ of neurons that share the same receptive field and have different response properties to image features. An app roximation of the optimal Bayesian observer, based on integrating local decision s, rather than diffusions, is also derived; it is shown experimentally to produc e very similar pre- dictions. A psychophyisics experiment is proposed that may d iscriminate between which mechanism is used in the human brain.
****
Sequence learning with hidden units in spiking neural networks
Johanni Brea, Walter Senn, Jean-pascal Pfister
We consider a statistical framework in which recurrent networks of spiking neuro ns learn to generate spatio-temporal spike patterns. Given biologically realisti c stochastic neuronal dynamics we derive a tractable learning rule for the synap tic weights towards hidden and visible neurons that leads to optimal recall of t he training sequences. We show that learning synaptic weights towards hidden neu rons significantly improves the storing capacity of the network. Furthermore, we derive an approximate online learning rule and show that our learning rule is c onsistent with Spike-Timing Dependent Plasticity in that if a presynaptic spike shortly precedes a postynaptic spike, potentiation is induced and otherwise depr ession is elicited.
****
Convergent Fitted Value Iteration with Linear Function Approximation
Daniel Lizotte
Fitted value iteration (FVI) with ordinary least squares regression is known to diverge. We present a new method, "Expansion-Constrained Ordinary Least Squares" (ECOLS), that produces a linear approximation but also guarantees convergence w hen used with FVI. To ensure convergence, we constrain the least squares regress

ion operator to be a non-expansion in the infinity-norm. We show that the space of function approximators that satisfy this constraint is more rich than the spa ce of "averagers," we prove a minimax property of the ECOLS residual error, and we give an efficient algorithm for computing the coefficients of ECOLS based on constraint generation. We illustrate the algorithmic convergence of FVI with ECO LS in a suite of experiments, and discuss its properties.
****

Learning in Hilbert vs. Banach Spaces: A Measure Embedding Viewpoint
Kenji Fukumizu, Gert Lanckriet, Bharath K. Sriperumbudur
The goal of this paper is to investigate the advantages and disadvantages of lea rning in Banach spaces over Hilbert spaces. While many works have been carried o ut in generalizing Hilbert methods to Banach spaces, in this paper, we consider the simple problem of learning a Parzen window classifier in a reproducing kerne l Banach space (RKBS)---which is closely related to the notion of embedding prob ability measures into an RKBS---in order to carefully understand its pros and co ns over the Hilbert space classifier. We show that while this generalization yie lds richer distance measures on probabilities compared to its Hilbert space coun terpart, it however suffers from serious computational drawback limiting its pra ctical applicability, which therefore demonstrates the need for developing effic ient learning algorithms in Banach spaces.
****

Predicting Dynamic Difficulty
Olana Missura, Thomas Gärtner
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.
****

Optimistic Optimization of a Deterministic Function without the Knowledge of its Smoothness
Rémi Munos
We consider a global optimization problem of a deterministic function f in a sem imetric space, given a finite budget of n evaluations. The function f is assumed to be locally smooth (around one of its global maxima) with respect to a semi-m etric . We describe two algorithms based on optimistic exploration that use a hi erarchical partitioning of the space at all scales. A first contribution is an a lgorithm, DOO, that requires the knowledge of . We report a finite-sample perfor mance bound in terms of a measure of the quantity of near-optimal states. We the n define a second algorithm, SOO, which does not require the knowledge of the se mimetric under which f is smooth, and whose performance is almost as good as DO O optimally-fitted.
****

Robust Multi-Class Gaussian Process Classification
Daniel Hernández-lobato, Jose Hernández-lobato, Pierre Dupont
Multi-class Gaussian Process Classifiers (MGPCs) are often affected by over-fitt ing problems when labeling errors occur far from the decision boundaries. To pre vent this, we investigate a robust MGPC (RMGPC) which considers labeling errors independently of their distance to the decision boundaries. Expectation propagat ion is used for approximate inference. Experiments with several datasets in whic h noise is injected in the class labels illustrate the benefits of RMGPC. This m ethod performs better than other Gaussian process alternatives based on consider ing latent Gaussian noise or heavy-tailed processes. When no noise is injected i n the labels, RMGPC still performs equal or better than the other methods. Final ly, we show how RMGPC can be used for successfully identifying data instances wh ich are difficult to classify accurately in practice.
****

Practical Variational Inference for Neural Networks
Alex Graves
Variational methods have been previously explored as a tractable approximation t o Bayesian inference for neural networks. However the approaches proposed so far

have only been applicable to a few simple network architectures. This paper introduces an easy-to-implement stochastic variational method (or equivalently, minimum description length loss function) that can be applied to most neural networks. Along the way it revisits several common regularisers from a variational perspective. It also provides a simple pruning heuristic that can both drastically reduce the number of network weights and lead to improved generalisation. Experimental results are provided for a hierarchical multidimensional recurrent neural network applied to the TIMIT speech corpus.
****

Penalty Decomposition Methods for Rank Minimization
Yong Zhang, Zhaosong Lu
In this paper we consider general rank minimization problems with rank appearing in either objective function or constraint. We first show that a class of matrix optimization problems can be solved as lower dimensional vector optimization problems. As a consequence, we establish that a class of rank minimization problems have closed form solutions. Using this result, we then propose penalty decomposition methods for general rank minimization problems. The convergence results of the PD methods have been shown in the longer version of the paper. Finally, we test the performance of our methods by applying them to matrix completion and nearest low-rank correlation matrix problems. The computational results demonstrate that our methods generally outperform the existing methods in terms of solution quality and/or speed.
****

Accelerated Adaptive Markov Chain for Partition Function Computation
Stefano Ermon, Carla P. Gomes, Ashish Sabharwal, Bart Selman
We propose a novel Adaptive Markov Chain Monte Carlo algorithm to compute the partition function. In particular, we show how to accelerate a flat histogram sampling technique by significantly reducing the number of ``null moves'' in the chain, while maintaining asymptotic convergence properties. Our experiments show that our method converges quickly to highly accurate solutions on a range of benchmark instances, outperforming other state-of-the-art methods such as IJGP, TRW, and Gibbs sampling both in run-time and accuracy. We also show how obtaining a so-called density of states distribution allows for efficient weight learning in Markov Logic theories.
****

On Strategy Stitching in Large Extensive Form Multiplayer Games
Richard Gibson, Duane Szafron
Computing a good strategy in a large extensive form game often demands an extraordinary amount of computer memory, necessitating the use of abstraction to reduce the game size. Typically, strategies from abstract games perform better in the real game as the granularity of abstraction is increased. This paper investigates two techniques for stitching a base strategy in a coarse abstraction of the full game tree, to expert strategies in fine abstractions of smaller subtrees. We provide a general framework for creating static experts, an approach that generalizes some previous strategy stitching efforts. In addition, we show that static experts can create strong agents for both 2-player and 3-player Leduc and Limit Texas Hold'em poker, and that a specific class of static experts can be preferred among a number of alternatives. Furthermore, we describe a poker agent that used static experts and won the 3-player events of the 2010 Annual Computer Poker Competition.
****

Active Learning Ranking from Pairwise Preferences with Almost Optimal Query Complexity
Nir Ailon
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.
****

Nonnegative dictionary learning in the exponential noise model for adaptive musi

c signal representation
Onur Dikmen, Cédric Févotte
In this paper we describe a maximum likelihood likelihood approach for dictionar y learning in the multiplicative exponential noise model. This model is prevalen t in audio signal processing where it underlies a generative composite model of the power spectrogram. Maximum joint likelihood estimation of the dictionary and expansion coefficients leads to a nonnegative matrix factorization problem wher e the Itakura-Saito divergence is used. The optimality of this approach is in qu estion because the number of parameters (which include the expansion coefficient s) grows with the number of observations. In this paper we describe a variationa l procedure for optimization of the marginal likelihood, i.e., the likelihood of the dictionary where the activation coefficients have been integrated out (give n a specific prior). We compare the output of both maximum joint likelihood esti mation (i.e., standard Itakura-Saito NMF) and maximum marginal likelihood estima tion (MMLE) on real and synthetical datasets. The MMLE approach is shown to embe d automatic model order selection, akin to automatic relevance determination.
****

From Stochastic Nonlinear Integrate-and-Fire to Generalized Linear Models
Skander Mensi, Richard Naud, Wulfram Gerstner
Variability in single neuron models is typically implemented either by a stochas tic Leaky-Integrate-and-Fire model or by a model of the Generalized Linear Model (GLM) family. We use analytical and numerical methods to relate state-of-the-ar t models from both schools of thought. First we find the analytical expressions relating the subthreshold voltage from the Adaptive Exponential Integrate-and-Fi re model (AdEx) to the Spike-Response Model with escape noise (SRM as an example of a GLM). Then we calculate numerically the link-function that provides the fi ring probability given a deterministic membrane potential. We find a mathematica l expression for this link-function and test the ability of the GLM to predict t he firing probability of a neuron receiving complex stimulation. Comparing the p rediction performance of various link-functions, we find that a GLM with an expo nential link-function provides an excellent approximation to the Adaptive Expone ntial Integrate-and-Fire with colored-noise input. These results help to underst and the relationship between the different approaches to stochastic neuron model s.
****

Generalised Coupled Tensor Factorisation
Kenan Y■lmaz, Ali Cemgil, Umut Simsekli
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.
****

Prismatic Algorithm for Discrete D.C. Programming Problem
Yoshinobu Kawahara, Takashi Washio
In this paper, we propose the first exact algorithm for minimizing the differenc e of two submodular functions (D.S.), i.e., the discrete version of the D.C. pro gramming problem. The developed algorithm is a branch-and-bound-based algorithm which responds to the structure of this problem through the relationship between submodularity and convexity. The D.S. programming problem covers a broad range of applications in machine learning because this generalizes the optimization of a wide class of set functions. We empirically investigate the performance of ou r algorithm, and illustrate the difference between exact and approximate solutio ns respectively obtained by the proposed and existing algorithms in feature sele ction and discriminative structure learning.
****

On the Completeness of First-Order Knowledge Compilation for Lifted Probabilisti c Inference
Guy Broeck
Probabilistic logics are receiving a lot of attention today because of their exp ressive power for knowledge representation and learning. However, this expressiv

ity is detrimental to the tractability of inference, when done at the propositio
nal level. To solve this problem, various lifted inference algorithms have been
 proposed that reason at the first-order level, about groups of objects as a who
le. Despite the existence of various lifted inference approaches, there are curr
ently no completeness results about these algorithms. The key contribution of t
his paper is that we introduce a formal definition of lifted inference that allo
ws us to reason about the completeness of lifted inference algorithms relative t
o a particular class of probabilistic models. We then show how to obtain a comp
leteness result using a first-order knowledge compilation approach for theories
of formulae containing up to two logical variables.
****

Analysis and Improvement of Policy Gradient Estimation
Tingting Zhao, Hirotaka Hachiya, Gang Niu, Masashi Sugiyama
Policy gradient is a useful model-free reinforcement learning approach, but it
tends to suffer from instability of gradient estimates. In this paper, we analy
ze and improve the stability of policy gradient methods. We first prove that the
 variance of gradient estimates in the PGPE(policy gradients with parameter-base
d exploration) method is smaller than that of the classical REINFORCE method und
er a mild assumption. We then derive the optimal baseline for PGPE, which contr
ibutes to further reducing the variance. We also theoretically show that PGPE w
ith the optimal baseline is more preferable than REINFORCE with the optimal base
line in terms of the variance of gradient estimates. Finally, we demonstrate th
e usefulness of the improved PGPE method through experiments.
****

On Tracking The Partition Function
Guillaume Desjardins, Yoshua Bengio, Aaron C. Courville
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.
****

Portmanteau Vocabularies for Multi-Cue Image Representation
Fahad Khan, Joost Weijer, Andrew Bagdanov, Maria Vanrell
We describe a novel technique for feature combination in the bag-of-words model
of image classification. Our approach builds discriminative compound words from
primitive cues learned independently from training images. Our main observation
is that modeling joint-cue distributions independently is more statistically rob
ust for typical classification problems than attempting to empirically estimate
the dependent, joint-cue distribution directly. We use Information theoretic voc
abulary compression to find discriminative combinations of cues and the resultin
g vocabulary of portmanteau words is compact, has the cue binding property, and
supports individual weighting of cues in the final image representation. State-o
f-the-art results on both the Oxford Flower-102 and Caltech-UCSD Bird-200 datase
ts demonstrate the effectiveness of our technique compared to other, significant
ly more complex approaches to multi-cue image representation
****

Algorithms for Hyper-Parameter Optimization
James Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl
Several recent advances to the state of the art in image classification benchmar
ks have come from better configurations of existing techniques rather than novel
 approaches to feature learning. Traditionally, hyper-parameter optimization has
 been the job of humans because they can be very efficient in regimes where only
 a few trials are possible. Presently, computer clusters and GPU processors make
 it possible to run more trials and we show that algorithmic approaches can find
 better results. We present hyper-parameter optimization results on tasks of tra
ining neural networks and deep belief networks (DBNs). We optimize hyper-paramet
ers using random search and two new greedy sequential methods based on the expec
ted improvement criterion. Random search has been shown to be sufficiently effic
ient for learning neural networks for several datasets, but we show it is unreli
able for training DBNs. The sequential algorithms are applied to the most diffic

ult DBN learning problems from [Larochelle et al., 2007] and find significantly better results than the best previously reported. This work contributes novel te chniques for making response surface models P (y|x) in which many elements of hy per-parameter assignment (x) are known to be irrelevant given particular values of other elements.

****

Finite Time Analysis of Stratified Sampling for Monte Carlo

Alexandra Carpentier, Rémi Munos

We consider the problem of stratified sampling for Monte-Carlo integration. We m odel this problem in a multi-armed bandit setting, where the arms represent the strata, and the goal is to estimate a weighted average of the mean values of the arms. We propose a strategy that samples the arms according to an upper bound o n their standard deviations and compare its estimation quality to an ideal alloc ation that would know the standard deviations of the arms. We provide two regret analyses: a distribution-dependent bound O(n^{-3/2}) that depends on a measure of the disparity of the arms, and a distribution-free bound O(n^{-4/3}) that doe s not. To the best of our knowledge, such a finite-time analysis is new for this problem.

****

Online Submodular Set Cover, Ranking, and Repeated Active Learning

Andrew Guillory, Jeff A. Bilmes

We propose an online prediction version of submodular set cover with connections to ranking and repeated active learning. In each round, the learning algorithm chooses a sequence of items. The algorithm then receives a monotone submodular function and suffers loss equal to the cover time of the function: the number o f items needed, when items are selected in order of the chosen sequence, to achi eve a coverage constraint. We develop an online learning algorithm whose loss c onverges to approximately that of the best sequence in hindsight. Our proposed algorithm is readily extended to a setting where multiple functions are revealed at each round and to bandit and contextual bandit settings.

****

The Fast Convergence of Boosting

Matus Telgarsky

This manuscript considers the convergence rate of boosting under a large class o f losses, including the exponential and logistic losses, where the best previous rate of convergence was $O(\exp(1/\varepsilon^2))$. First, it is established that the settin g of weak learnability aids the entire class, granting a rate $O(\ln(1/\varepsilon))$. Next, the (disjoint) conditions under which the infimal empirical risk is attainable are characterized in terms of the sample and weak learning class, and a new proo f is given for the known rate $O(\ln(1/\varepsilon))$. Finally, it is established that any i nstance can be decomposed into two smaller instances resembling the two precedin g special cases, yielding a rate $O(1/\varepsilon)$, with a matching lower bound for the log istic loss. The principal technical hurdle throughout this work is the potentia l unattainability of the infimal empirical risk; the technique for overcoming th is barrier may be of general interest.

****

See the Tree Through the Lines: The Shazoo Algorithm

Fabio Vitale, Nicolò Cesa-bianchi, Claudio Gentile, Giovanni Zappella

Predicting the nodes of a given graph is a fascinating theoretical problem wit h applications in several domains. Since graph sparsification via spanning tre es retains enough information while making the task much easier, trees are a n important special case of this problem. Although it is known how to predict the nodes of an unweighted tree in a nearly optimal way, in the weighted case a fully satisfactory algorithm is not available yet. We fill this hole and int roduce an efficient node predictor, Shazoo, which is nearly optimal on any wei ghted tree. Moreover, we show that Shazoo can be viewed as a common nontrivial generalization of both previous approaches for unweighted trees and weighted lines. Experiments on real-world datasets confirm that Shazoo performs well in that it fully exploits the structure of the input tree, and gets very close to (and sometimes better than) less scalable energy minimization methods.

****
t-divergence Based Approximate Inference
Nan Ding, Yuan Qi, S.v.n. Vishwanathan
Approximate inference is an important technique for dealing with large, intractable graphical models based on the exponential family of distributions. We extend the idea of approximate inference to the t-exponential family by defining a new t-divergence. This divergence measure is obtained via convex duality between the log-partition function of the t-exponential family and a new t-entropy. We illustrate our approach on the Bayes Point Machine with a Student's t-prior.
****
Boosting with Maximum Adaptive Sampling
Charles Dubout, Francois Fleuret
Classical Boosting algorithms, such as AdaBoost, build a strong classifier without concern about the computational cost. Some applications, in particular in computer vision, may involve up to millions of training examples and features. In such contexts, the training time may become prohibitive. Several methods exist to accelerate training, typically either by sampling the features, or the examples, used to train the weak learners. Even if those methods can precisely quantify the speed improvement they deliver, they offer no guarantee of being more efficient than any other, given the same amount of time. This paper aims at shading some light on this problem, i.e. given a fixed amount of time, for a particular problem, which strategy is optimal in order to reduce the training loss the most. We apply this analysis to the design of new algorithms which estimate on the fly at every iteration the optimal trade-off between the number of samples and the number of features to look at in order to maximize the expected loss reduction. Experiments in object recognition with two standard computer vision data-sets show that the adaptive methods we propose outperform basic sampling and state-of-the-art bandit methods.
****
Inverting Grice's Maxims to Learn Rules from Natural Language Extractions
Mohammad Sorower, Janardhan Doppa, Walker Orr, Prasad Tadepalli, Thomas Dietterich, Xiaoli Fern
We consider the problem of learning rules from natural language text sources. These sources, such as news articles and web texts, are created by a writer to communicate information to a reader, where the writer and reader share substantial domain knowledge. Consequently, the texts tend to be concise and mention the minimum information necessary for the reader to draw the correct conclusions. We study the problem of learning domain knowledge from such concise texts, which is an instance of the general problem of learning in the presence of missing data. However, unlike standard approaches to missing data, in this setting we know that facts are more likely to be missing from the text in cases where the reader can infer them from the facts that are mentioned combined with the domain knowledge. Hence, we can explicitly model this "missingness" process and invert it via probabilistic inference to learn the underlying domain knowledge. This paper introduces a mention model that models the probability of facts being mentioned in the text based on what other facts have already been mentioned and domain knowledge in the form of Horn clause rules. Learning must simultaneously search the space of rules and learn the parameters of the mention model. We accomplish this via an application of Expectation Maximization within a Markov Logic framework. An experimental evaluation on synthetic and natural text data shows that the method can learn accurate rules and apply them to new texts to make correct inferences. Experiments also show that the method out-performs the standard EM approach that assumes mentions are missing at random.
****
Efficient anomaly detection using bipartite k-NN graphs
Kumar Sricharan, Alfred Hero
Learning minimum volume sets of an underlying nominal distribution is a very effective approach to anomaly detection. Several approaches to learning minimum volume sets have been proposed in the literature, including the K-point nearest neighbor graph (K-kNNG) algorithm based on the geometric entropy minimization (GEM)

principle [4]. The K-kNNG detector, while possessing several desirable characte
ristics, suffers from high computation complexity, and in [4] a simpler heuristi
c approximation, the leave-one-out kNNG (L1O-kNNG) was proposed. In this paper,
we propose a novel bipartite k-nearest neighbor graph (BP-kNNG) anomaly detectio
n scheme for estimating minimum volume sets. Our bipartite estimator retains all
the desirable theoretical properties of the K-kNNG, while being computationally
simpler than the K-kNNG and the surrogate L1O-kNNG detectors. We show that BP-k
NNG is asymptotically consistent in recovering the p-value of each test point. E
xperimental results are given that illustrate the superior performance of BP-kNN
G as compared to the L1O-kNNG and other state of the art anomaly detection schem
es.
****

Shallow vs. Deep Sum-Product Networks
Olivier Delalleau, Yoshua Bengio
We investigate the representational power of sum-product networks (computation
networks analogous to neural networks, but whose individual units compute eithe
r products or weighted sums), through a theoretical analysis that compares dee
p (multiple hidden layers) vs. shallow (one hidden layer) architectures. We pro
ve there exist families of functions that can be represented much more efficien
tly with a deep network than with a shallow one, i.e. with substantially fewer
hidden units. Such results were not available until now, and contribute to mot
ivate recent research involving learning of deep sum-product networks, and more
generally motivate research in Deep Learning.
****

Expressive Power and Approximation Errors of Restricted Boltzmann Machines
Guido F. Montufar, Johannes Rauh, Nihat Ay
We present explicit classes of probability distributions that can be learned by
Restricted Boltzmann Machines (RBMs) depending on the number of units that they
contain, and which are representative for the expressive power of the model. We
use this to show that the maximal Kullback-Leibler divergence to the RBM model w
ith n visible and m hidden units is bounded from above by (n-1)-log(m+1). In thi
s way we can specify the number of hidden units that guarantees a sufficiently r
ich model containing different classes of distributions and respecting a given e
rror tolerance.
****

Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-t
ype Operators
Dominique Perrault-joncas, Marina Meila
This paper considers the problem of embedding directed graphs in Euclidean space
while retaining directional information. We model the observed graph as a sampl
e from a manifold endowed with a vector field, and we design an algo- rithm that
separates and recovers the features of this process: the geometry of the manifo
ld, the data density and the vector field. The algorithm is motivated by our ana
lysis of Laplacian-type operators and their continuous limit as generators of di
ffusions on a manifold. We illustrate the recovery algorithm on both artificiall
y constructed and real data.
****

Information Rates and Optimal Decoding in Large Neural Populations
Kamiar Rad, Liam Paninski
Many fundamental questions in theoretical neuroscience involve optimal decoding
and the computation of Shannon information rates in populations of spiking neu
rons. In this paper, we apply methods from the asymptotic theory of statistica
l inference to obtain a clearer analytical understanding of these quantities.
We find that for large neural populations carrying a finite total amount of inf
ormation, the full spiking population response is asymptotically as informative
as a single observation from a Gaussian process whose mean and covariance can
be characterized explicitly in terms of network and single neuron properties.
The Gaussian form of this asymptotic sufficient statistic allows us in certain
cases to perform optimal Bayesian decoding by simple linear transformations, a
nd to obtain closed-form expressions of the Shannon information carried by the

network. One technical advantage of the theory is that it may be applied easily even to non-Poisson point process network models; for example, we find that under some conditions, neural populations with strong history-dependent (non-Poisson) effects carry exactly the same information as do simpler equivalent populations of non-interacting Poisson neurons with matched firing rates. We argue that our findings help to clarify some results from the recent literature on neural decoding and neuroprosthetic design.
****

Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization
Mark Schmidt, Nicolas Roux, Francis Bach
We consider the problem of optimizing the sum of a smooth convex function and a non-smooth convex function using proximal-gradient methods, where an error is present in the calculation of the gradient of the smooth term or in the proximity operator with respect to the second term. We show that the basic proximal-gradient method, the basic proximal-gradient method with a strong convexity assumption, and the accelerated proximal-gradient method achieve the same convergence rates as in the error-free case, provided the errors decrease at an appropriate rate. Our experimental results on a structured sparsity problem indicate that sequences of errors with these appealing theoretical properties can lead to practical performance improvements.
****

Co-Training for Domain Adaptation
Minmin Chen, Kilian Q. Weinberger, John Blitzer
Domain adaptation algorithms seek to generalize a model trained in a source domain to a new target domain. In many practical cases, the source and target distributions can differ substantially, and in some cases crucial target features may not have support in the source domain. In this paper we introduce an algorithm that bridges the gap between source and target domains by slowly adding both the target features and instances in which the current algorithm is the most confident. Our algorithm is a variant of co-training, and we name it CODA (Co-training for domain adaptation). Unlike the original co-training work, we do not assume a particular feature split. Instead, for each iteration of co-training, we add target features and formulate a single optimization problem which simultaneously learns a target predictor, a split of the feature space into views, and a shared subset of source and target features to include in the predictor. CODA significantly out-performs the state-of-the-art on the 12-domain benchmark data set of Blitzer et al.. Indeed, over a wide range (65 of 84 comparisons) of target supervision, ranging from no labeled target data to a relatively large number of target labels, CODA achieves the best performance.
****

Learning to Agglomerate Superpixel Hierarchies
Viren Jain, Srinivas C. Turaga, K Briggman, Moritz Helmstaedter, Winfried Denk, H. Seung
An agglomerative clustering algorithm merges the most similar pair of clusters at every iteration. The function that evaluates similarity is traditionally hand-designed, but there has been recent interest in supervised or semisupervised settings in which ground-truth clustered data is available for training. Here we show how to train a similarity function by regarding it as the action-value function of a reinforcement learning problem. We apply this general method to segment images by clustering superpixels, an application that we call Learning to Agglomerate Superpixel Hierarchies (LASH). When applied to a challenging dataset of brain images from serial electron microscopy, LASH dramatically improved segmentation accuracy when clustering supervoxels generated by state of the boundary detection algorithms. The naive strategy of directly training only supervoxel similarities and applying single linkage clustering produced less improvement.
****

Structured Learning for Cell Tracking
Xinghua Lou, Fred A. Hamprecht
We study the problem of learning to track a large quantity of homogeneous objects such as cell tracking in cell culture study and developmental biology. Reliabl

e cell tracking in time-lapse microscopic image sequences is important for moder n biomedical research. Existing cell tracking methods are usually kept simple an d use only a small number of features to allow for manual parameter tweaking or grid search. We propose a structured learning approach that allows to learn opti mum parameters automatically from a training set. This allows for the use of a r icher set of features which in turn affords improved tracking compared to recent ly reported methods on two public benchmark sequences.
****

Hierarchical Topic Modeling for Analysis of Time-Evolving Personal Choices
Xianxing Zhang, Lawrence Carin, David Dunson
The nested Chinese restaurant process is extended to design a nonparametric top ic-model tree for representation of human choices. Each tree branch corresponds to a type of person, and each node (topic) has a corresponding probability vec tor over items that may be selected. The observed data are assumed to have asso ciated temporal covariates (corresponding to the time at which choices are made), and we wish to impose that with increasing time it is more probable that top ics deeper in the tree are utilized. This structure is imposed by developing a new “change point" stick-breaking model that is coupled with a Poisson and prod uct-of-gammas construction. To share topics across the tree nodes, topic distrib utions are drawn from a Dirichlet process. As a demonstration of this concept, we analyze real data on course selections of undergraduate students at Duke Uni versity, with the goal of uncovering and concisely representing structure in th e curriculum and in the characteristics of the student body.
****

Selecting the State-Representation in Reinforcement Learning
Odalric-ambrym Maillard, Daniil Ryabko, Rémi Munos
The problem of selecting the right state-representation in a reinforcement learn ing problem is considered. Several models (functions mapping past observations to a finite set) of the observations are given, and it is known that for at l east one of these models the resulting state dynamics are indeed Markovian. With out knowing neither which of the models is the correct one, nor what are the pr obabilistic characteristics of the resulting MDP, it is required to obtain as much reward as the optimal policy for the correct model (or for the best of the correct models, if there are several). We propose an algorithm that achieves tha t, with a regret of order T^{2/3} where T is the horizon time.
****

A Reinforcement Learning Theory for Homeostatic Regulation
Mehdi Keramati, Boris Gutkin
Reinforcement learning models address animal's behavioral adaptation to its chan ging "external" environment, and are based on the assumption that Pavlovian, hab itual and goal-directed responses seek to maximize reward acquisition. Negative- feedback models of homeostatic regulation, on the other hand, are concerned with behavioral adaptation in response to the "internal" state of the animal, and as sume that animals' behavioral objective is to minimize deviations of some key ph ysiological variables from their hypothetical setpoints. Building upon the drive -reduction theory of reward, we propose a new analytical framework that integrat es learning and regulatory systems, such that the two seemingly unrelated object ives of reward maximization and physiological-stability prove to be identical. T he proposed theory shows behavioral adaptation to both internal and external sta tes in a disciplined way. We further show that the proposed framework allows for a unified explanation of some behavioral phenomenon like motivational sensitivi ty of different associative learning mechanism, anticipatory responses, interact ion among competing motivational systems, and risk aversion.
****

Semantic Labeling of 3D Point Clouds for Indoor Scenes
Hema Koppula, Abhishek Anand, Thorsten Joachims, Ashutosh Saxena
Inexpensive RGB-D cameras that give an RGB image together with depth data have b ecome widely available. In this paper, we use this data to build 3D point clouds of full indoor scenes such as an office and address the task of semantic labeli ng of these 3D point clouds. We propose a graphical model that captures various

features and contextual relations, including the local visual appearance and sha pe cues, object co-occurence relationships and geometric relationships. With a l arge number of object classes and relations, the model’s parsimony becomes impor tant and we address that by using multiple types of edge potentials. The model a dmits efficient approximate inference, and we train it using a maximum-margin le arning approach. In our experiments over a total of 52 3D scenes of homes and of fices (composed from about 550 views, having 2495 segments labeled with 27 objec t classes), we get a performance of 84.06% in labeling 17 object classes for off ices, and 73.38% in labeling 17 object classes for home scenes. Finally, we appl ied these algorithms successfully on a mobile robot for the task of finding obje cts in large cluttered rooms.
****

Higher-Order Correlation Clustering for Image Segmentation
Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, Chang Yoo
For many of the state-of-the-art computer vision algorithms, image segmentation is an important preprocessing step. As such, several image segmentation algorith ms have been proposed, however, with certain reservation due to high computation al load and many hand-tuning parameters. Correlation clustering, a graph-partiti oning algorithm often used in natural language processing and document clusterin g, has the potential to perform better than previously proposed image segmentati on algorithms. We improve the basic correlation clustering formulation by taking into account higher-order cluster relationships. This improves clustering in th e presence of local boundary ambiguities. We first apply the pairwise correlatio n clustering to image segmentation over a pairwise superpixel graph and then dev elop higher-order correlation clustering over a hypergraph that considers higher -order relations among superpixels. Fast inference is possible by linear program ming relaxation, and also effective parameter learning framework by structured s upport vector machine is possible. Experimental results on various datasets show that the proposed higher-order correlation clustering outperforms other state-o f-the-art image segmentation algorithms.
****

Learning large-margin halfspaces with more malicious noise
Phil Long, Rocco Servedio
We describe a simple algorithm that runs in time poly(n,1/gamma,1/eps) and lear ns an unknown n-dimensional gamma-margin halfspace to accuracy 1-eps in the pre sence of malicious noise, when the noise rate is allowed to be as high as Thet a(eps gamma sqrt(log(1/gamma))). Previous efficient algorithms could only learn to accuracy eps in the presence of malicious noise of rate at most Theta(eps g amma). Our algorithm does not work by optimizing a convex loss function. We show that no algorithm for learning gamma-margin halfspaces that minimizes a c onvex proxy for misclassification error can tolerate malicious noise at a rate greater than Theta(eps gamma); this may partially explain why previous algorith ms could not achieve the higher noise tolerance of our new algorithm.
****

The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning
Marius Kloft, Gilles Blanchard
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 Global Structural EM Algorithm for a Model of Cancer Progression
Ali Tofigh, Erik Sj■lund, Mattias H■glund, Jens Lagergren
Cancer has complex patterns of progression that include converging as well as d iverging progressional pathways. Vogelstein's path model of colon cancer was a pioneering contribution to cancer research. Since then, several attempts have been made at obtaining mathematical models of cancer progression, devising lear ning algorithms, and applying these to cross-sectional data. Beerenwinkel {\em et al.} provided, what they coined, EM-like algorithms for Oncogenetic Trees (OTs) and mixtures of such. Given the small size of current and future data set

s, it is important to minimize the number of parameters of a model. For this reason, we too focus on tree-based models and introduce Hidden-variable Oncogenetic Trees (HOTs). In contrast to OTs, HOTs allow for errors in the data and thereby provide more realistic modeling. We also design global structural EM algorithms for learning HOTs and mixtures of HOTs (HOT-mixtures). The algorithms are global in the sense that, during the M-step, they find a structure that yields a global maximum of the expected complete log-likelihood rather than merely one that improves it. The algorithm for single HOTs performs very well on reasonable-sized data sets, while that for HOT-mixtures requires data sets of sizes obtainable only with tomorrow's more cost-efficient technologies.
****

Infinite Latent SVM for Classification and Multi-task Learning
Jun Zhu, Ning Chen, Eric Xing
Unlike existing nonparametric Bayesian models, which rely solely on specially conceived priors to incorporate domain knowledge for discovering improved latent representations, we study nonparametric Bayesian inference with regularization on the desired posterior distributions. While priors can indirectly affect posterior distributions through Bayes' theorem, imposing posterior regularization is arguably more direct and in some cases can be much easier. We particularly focus on developing infinite latent support vector machines (iLSVM) and multi-task infinite latent support vector machines (MT-iLSVM), which explore the large-margin idea in combination with a nonparametric Bayesian model for discovering predictive latent features for classification and multi-task learning, respectively. We present efficient inference methods and report empirical studies on several benchmark datasets. Our results appear to demonstrate the merits inherited from both large-margin learning and Bayesian nonparametrics.
****

On U-processes and clustering performance
Stéphan Clémençcon
Many clustering techniques aim at optimizing empirical criteria that are of the form of a U-statistic of degree two. Given a measure of dissimilarity between pairs of observations, the goal is to minimize the within cluster point scatter over a class of partitions of the feature space. It is the purpose of this paper to define a general statistical framework, relying on the theory of U-processes, for studying the performance of such clustering methods. In this setup, under adequate assumptions on the complexity of the subsets forming the partition candidates, the excess of clustering risk is proved to be of the order O(1/\sqrt{n}). Based on recent results related to the tail behavior of degenerate U-processes, it is also shown how to establish tighter rate bounds. Model selection issues, related to the number of clusters forming the data partition in particular, are also considered.
****

Robust Lasso with missing and grossly corrupted observations
Nasser Nasrabadi, Trac Tran, Nam Nguyen
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.
****

Unifying Non-Maximum Likelihood Learning Objectives with Minimum KL Contraction
Siwei Lyu
When used to learn high dimensional parametric probabilistic models, the clas- sical maximum likelihood (ML) learning often suffers from computational in- tractability, which motivates the active developments of non-ML learning meth- ods. Yet, because of their divergent motivations and forms, the objective func- tions of many non-ML learning methods are seemingly unrelated, and there lacks a unified framework to understand them. In this work, based on an information geometric view of parametric learning, we introduce a general non-ML learning principle termed as minimum KL contraction, where we seek optimal parameters that minimizes the contraction of the KL divergence between the two distributions after they a

re transformed with a KL contraction operator. We then show that the objective f unctions of several important or recently developed non-ML learn- ing methods, i ncluding contrastive divergence [12], noise-contrastive estimation [11], partial likelihood [7], non-local contrastive objectives [31], score match- ing [14], p seudo-likelihood [3], maximum conditional likelihood [17], maximum mutual inform ation [2], maximum marginal likelihood [9], and conditional and marginal composi te likelihood [24], can be unified under the minimum KL con- traction framework with different choices of the KL contraction operators.
****
Select and Sample - A Model of Efficient Neural Inference and Learning
Jacquelyn Shelton, Abdul Sheikh, Pietro Berkes, Joerg Bornschein, Jörg Lücke
An increasing number of experimental studies indicate that perception encodes a posterior probability distribution over possible causes of sensory stimuli, whic h is used to act close to optimally in the environment. One outstanding difficul ty with this hypothesis is that the exact posterior will in general be too compl ex to be represented directly, and thus neurons will have to represent an approx imation of this distribution. Two influential proposals of efficient posterior r epresentation by neural populations are: 1) neural activity represents samples o f the underlying distribution, or 2) they represent a parametric representation of a variational approximation of the posterior. We show that these approaches c an be combined for an inference scheme that retains the advantages of both: it i s able to represent multiple modes and arbitrary correlations, a feature of samp ling methods, and it reduces the represented space to regions of high probabilit y mass, a strength of variational approximations. Neurally, the combined method can be interpreted as a feed-forward preselection of the relevant state space, f ollowed by a neural dynamics implementation of Markov Chain Monte Carlo (MCMC) t o approximate the posterior over the relevant states. We demonstrate the effecti veness and efficiency of this approach on a sparse coding model. In numerical ex periments on artificial data and image patches, we compare the performance of th e algorithms to that of exact EM, variational state space selection alone, MCMC alone, and the combined select and sample approach. The select and sample approa ch integrates the advantages of the sampling and variational approximations, and forms a robust, neurally plausible, and very efficient model of processing and learning in cortical networks. For sparse coding we show applications easily exc eeding a thousand observed and a thousand hidden dimensions.
****
Clustered Multi-Task Learning Via Alternating Structure Optimization
Jiayu Zhou, Jianhui Chen, Jieping Ye
Multi-task learning (MTL) learns multiple related tasks simultaneously to improv e generalization performance. Alternating structure optimization (ASO) is a popu lar MTL method that learns a shared low-dimensional predictive structure on hypo thesis spaces from multiple related tasks. It has been applied successfully in m any real world applications. As an alternative MTL approach, clustered multi-tas k learning (CMTL) assumes that multiple tasks follow a clustered structure, i.e. , tasks are partitioned into a set of groups where tasks in the same group are s imilar to each other, and that such a clustered structure is unknown a priori. T he objectives in ASO and CMTL differ in how multiple tasks are related. Interest ingly, we show in this paper the equivalence relationship between ASO and CMTL, providing significant new insights into ASO and CMTL as well as their inherent r elationship. The CMTL formulation is non-convex, and we adopt a convex relaxatio n to the CMTL formulation. We further establish the equivalence relationship bet ween the proposed convex relaxation of CMTL and an existing convex relaxation of ASO, and show that the proposed convex CMTL formulation is significantly more e fficient especially for high-dimensional data. In addition, we present three alg orithms for solving the convex CMTL formulation. We report experimental results on benchmark datasets to demonstrate the efficiency of the proposed algorithms.
****
On Learning Discrete Graphical Models using Greedy Methods
Ali Jalali, Christopher Johnson, Pradeep Ravikumar
In this paper, we address the problem of learning the structure of a pairwise gr

aphical model from samples in a high-dimensional setting. Our first main result studies the sparsistency, or consistency in sparsity pattern recovery, properties of a forward-backward greedy algorithm as applied to general statistical models. As a special case, we then apply this algorithm to learn the structure of a discrete graphical model via neighborhood estimation. As a corollary of our general result, we derive sufficient conditions on the number of samples n, the maximum node-degree d and the problem size p, as well as other conditions on the model parameters, so that the algorithm recovers all the edges with high probability. Our result guarantees graph selection for samples scaling as n = Omega(d log(p)), in contrast to existing convex-optimization based algorithms that require a sample complexity of Omega(d^2 log(p)). Further, the greedy algorithm only requires a restricted strong convexity condition which is typically milder than irrepresentability assumptions. We corroborate these results using numerical simulations at the end.

****

Learning person-object interactions for action recognition in still images

Vincent Delaitre, Josef Sivic, Ivan Laptev

We investigate a discriminatively trained model of person-object interactions for recognizing common human actions in still images. We build on the locally order-less spatial pyramid bag-of-features model, which was shown to perform extremely well on a range of object, scene and human action recognition tasks. We introduce three principal contributions. First, we replace the standard quantized local HOG/SIFT features with stronger discriminatively trained body part and object detectors. Second, we introduce new person-object interaction features based on spatial co-occurrences of individual body parts and objects. Third, we address the combinatorial problem of a large number of possible interaction pairs and propose a discriminative selection procedure using a linear support vector machine (SVM) with a sparsity inducing regularizer. Learning of action-specific body part and object interactions bypasses the difficult problem of estimating the complete human body pose configuration. Benefits of the proposed model are shown on human action recognition in consumer photographs, outperforming the strong bag-of-features baseline.

****

Efficient inference in matrix-variate Gaussian models with \iid observation noise

Oliver Stegle, Christoph Lippert, Joris M. Mooij, Neil Lawrence, Karsten Borgwardt

Inference in matrix-variate Gaussian models has major applications for multi- output prediction and joint learning of row and column covariances from matrix- variate data. Here, we discuss an approach for ef■cient inference in such models that explicitly account for iid observation noise. Computational tractability can be retained by exploiting the Kronecker product between row and column covariance matrices. Using this framework, we show how to generalize the Graphical Lasso in order to learn a sparse inverse covariance between features while accounting for a low-rank confounding covariance between samples. We show practical utility on applications to biology, where we model covariances with more than 100,000 di- mensions. We ■nd greater accuracy in recovering biological network structures and are able to better reconstruct the confounders.

****

Confidence Sets for Network Structure

David Choi, Patrick Wolfe, Edo M. Airoldi

Latent variable models are frequently used to identify structure in dichotomous network data, in part because they give rise to a Bernoulli product likelihood that is both well understood and consistent with the notion of exchangeable random graphs. In this article we propose conservative confidence sets that hold with respect to these underlying Bernoulli parameters as a function of any given partition of network nodes, enabling us to assess estimates of \emph{residual} network structure, that is, structure that cannot be explained by known covariates and thus cannot be easily verified by manual inspection. We demonstrate the proposed methodology by analyzing student friendship networks from the National Long

itudinal Survey of Adolescent Health that include race, gender, and school year as covariates. We employ a stochastic expectation-maximization algorithm to fit a logistic regression model that includes these explanatory variables as well as a latent stochastic blockmodel component and additional node-specific effects. Although maximum-likelihood estimates do not appear consistent in this context, we are able to evaluate confidence sets as a function of different blockmodel partitions, which enables us to qualitatively assess the significance of estimated residual network structure relative to a baseline, which models covariates but lacks block structure.

****

Structural equations and divisive normalization for energy-dependent component analysis

Jun-ichiro Hirayama, Aapo Hyvärinen

Components estimated by independent component analysis and related methods are typically not independent in real data. A very common form of nonlinear dependency between the components is correlations in their variances or ener- gies. Here, we propose a principled probabilistic model to model the energy- correlations between the latent variables. Our two-stage model includes a linear mixing of latent signals into the observed ones like in ICA. The main new fea- ture is a model of the energy-correlations based on the structural equation model (SEM), in particular, a Linear Non-Gaussian SEM. The SEM is closely related to divisive normalization which effectively reduces energy correlation. Our new two- stage model enables estimation of both the linear mixing and the interactions re- lated to energy-correlations, without resorting to approximations of the likelihood function or other non-principled approaches. We demonstrate the applicability of our method with synthetic dataset, natural images and brain signals.

****

Gaussian Process Training with Input Noise

Andrew Mchutchon, Carl Rasmussen

In standard Gaussian Process regression input locations are assumed to be noise free. We present a simple yet effective GP model for training on input points corrupted by i.i.d. Gaussian noise. To make computations tractable we use a local linear expansion about each input point. This allows the input noise to be recast as output noise proportional to the squared gradient of the GP posterior mean. The input noise variances are inferred from the data as extra hyperparameters. They are trained alongside other hyperparameters by the usual method of maximisation of the marginal likelihood. Training uses an iterative scheme, which alternates between optimising the hyperparameters and calculating the posterior gradient. Analytic predictive moments can then be found for Gaussian distributed test points. We compare our model to others over a range of different regression problems and show that it improves over current methods.

****

On the Universality of Online Mirror Descent

Nati Srebro, Karthik Sridharan, Ambuj Tewari

We show that for a general class of convex online learning problems, Mirror Descent can always achieve a (nearly) optimal regret guarantee.

****

Unifying Framework for Fast Learning Rate of Non-Sparse Multiple Kernel Learning

Taiji Suzuki

In this paper, we give a new generalization error bound of Multiple Kernel Learning (MKL) for a general class of regularizations. Our main target in this paper is dense type regularizations including ■p-MKL that imposes ■p-mixed-norm regularization instead of ■1-mixed-norm regularization. According to the recent numerical experiments, the sparse regularization does not necessarily show a good performance compared with dense type regularizations. Motivated by this fact, this paper gives a general theoretical tool to derive fast learning rates that is applicable to arbitrary monotone norm-type regularizations in a unifying manner. As a by-product of our general result, we show a fast learning rate of ■p-MKL that is tightest among existing bounds. We also show that our general learning rate achieves the minimax lower bound. Finally, we show that, when the compl

exities of candidate reproducing kernel Hilbert spaces are inhomogeneous, dense type regularization shows better learning rate compared with sparse ■1 regularization.
****

Speedy Q-Learning
Mohammad Ghavamzadeh, Hilbert Kappen, Mohammad Azar, Rémi Munos
We introduce a new convergent variant of Q-learning, called speedy Q-learning, to address the problem of slow convergence in the standard form of the Q-learning algorithm. We prove a PAC bound on the performance of SQL, which shows that for an MDP with n state-action pairs and the discount factor \gamma only T=O\big(\log(n)/(\epsilon^{2}(1-\gamma)^{4})\big) steps are required for the SQL algorithm to converge to an \epsilon-optimal action-value function with high probability. This bound has a better dependency on 1/\epsilon and 1/(1-\gamma), and thus, is tighter than the best available result for Q-learning. Our bound is also superior to the existing results for both model-free and model-based instances of batch Q-value iteration that are considered to be more efficient than the incremental methods like Q-learning.
****

High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity
Po-ling Loh, Martin J. Wainwright
Although the standard formulations of prediction problems involve fully-observed and noiseless data drawn in an i.i.d. manner, many applications involve noisy and/or missing data, possibly involving dependencies. We study these issues in the context of high-dimensional sparse linear regression, and propose novel estimators for the cases of noisy, missing, and/or dependent data. Many standard approaches to noisy or missing data, such as those using the EM algorithm, lead to optimization problems that are inherently non-convex, and it is difficult to establish theoretical guarantees on practical algorithms. While our approach also involves optimizing non-convex programs, we are able to both analyze the statistical error associated with any global optimum, and prove that a simple projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of global minimizers. On the statistical side, we provide non-asymptotic bounds that hold with high probability for the cases of noisy, missing, and/or dependent data. On the computational side, we prove that under the same types of conditions required for statistical consistency, the projected gradient descent algorithm will converge at geometric rates to a near-global minimizer. We illustrate these theoretical predictions with simulations, showing agreement with the predicted scalings.
****

Greedy Model Averaging
Dong Dai, Tong Zhang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation
Nico Goernitz, Christian Widmer, Georg Zeller, Andre Kahles, Gunnar Rätsch, Sören Sonnenburg
We present a novel regularization-based Multitask Learning (MTL) formulation for Structured Output (SO) prediction for the case of hierarchical task relations. Structured output learning often results in dif■cult inference problems and requires large amounts of training data to obtain accurate models. We propose to use MTL to exploit information available for related structured output learning tasks by means of hierarchical regularization. Due to the combination of example sets, the cost of training models for structured output prediction can easily become infeasible for real world applications. We thus propose an ef■cient algorithm based on bundle methods to solve the optimization problems resulting

from MTL structured output learning. We demonstrate the performance of our approach on gene ■nding problems from the application domain of computational biology. We show that 1) our proposed solver achieves much faster convergence than previous methods and 2) that the Hierarchical SO-MTL approach clearly outperforms considered non-MTL methods.

****

Generalized Beta Mixtures of Gaussians

Artin Armagan, Merlise Clyde, David Dunson

In recent years, a rich variety of shrinkage priors have been proposed that have great promise in addressing massive regression problems. In general, these new priors can be expressed as scale mixtures of normals, but have more complex forms and better properties than traditional Cauchy and double exponential priors. We first propose a new class of normal scale mixtures through a novel generalized beta distribution that encompasses many interesting priors as special cases. This encompassing framework should prove useful in comparing competing priors, considering properties and revealing close connections. We then develop a class of variational Bayes approximations through the new hierarchy presented that will scale more efficiently to the types of truly massive data sets that are now encountered routinely.

****

Variational Gaussian Process Dynamical Systems

Andreas Damianou, Michalis Titsias, Neil Lawrence

High dimensional time series are endemic in applications of machine learning such as robotics (sensor data), computational biology (gene expression data), vision (video sequences) and graphics (motion capture data). Practical nonlinear probabilistic approaches to this data are required. In this paper we introduce the variational Gaussian process dynamical system. Our work builds on recent variational approximations for Gaussian process latent variable models to allow for nonlinear dimensionality reduction simultaneously with learning a dynamical prior in the latent space. The approach also allows for the appropriate dimensionality of the latent space to be automatically determined. We demonstrate the model on a human motion capture data set and a series of high resolution video sequences.

****

Statistical Tests for Optimization Efficiency

Levi Boyles, Anoop Korattikara, Deva Ramanan, Max Welling

Learning problems such as logistic regression are typically formulated as pure optimization problems defined on some loss function. We argue that this view ignores the fact that the loss function depends on stochastically generated data which in turn determines an intrinsic scale of precision for statistical estimation. By considering the statistical properties of the update variables used during the optimization (e.g. gradients), we can construct frequentist hypothesis tests to determine the reliability of these updates. We utilize subsets of the data for computing updates, and use the hypothesis tests for determining when the batch-size needs to be increased. This provides computational benefits and avoids overfitting by stopping when the batch-size has become equal to size of the full dataset. Moreover, the proposed algorithms depend on a single interpretable parameter – the probability for an update to be in the wrong direction – which is set to a single value across all algorithms and datasets. In this paper, we illustrate these ideas on three L1 regularized coordinate algorithms: L1 -regularized L2 -loss SVMs, L1 -regularized logistic regression, and the Lasso, but we emphasize that the underlying methods are much more generally applicable.

****

Statistical Performance of Convex Tensor Decomposition

Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, Hisashi Kashima

We analyze the statistical performance of a recently proposed convex tensor decomposition algorithm. Conventionally tensor decomposition has been formulated as non-convex optimization problems, which hindered the analysis of their performance. We show under some conditions that the mean squared error of the convex meth

od scales linearly with the quantity we call the normalized rank of the true ten
sor. The current analysis naturally extends the analysis of convex low-rank matr
ix estimation to tensors. Furthermore, we show through numerical experiments tha
t our theory can precisely predict the scaling behaviour in practice.
****

A Machine Learning Approach to Predict Chemical Reactions
Matthew Kayala, Pierre Baldi
Being able to predict the course of arbitrary chemical reactions is essential to
 the theory and applications of organic chemistry. Previous approaches are not h
igh-throughput, are not generalizable or scalable, or lack sufficient data to be
 effective. We describe single mechanistic reactions as concerted electron movem
ents from an electron orbital source to an electron orbital sink. We use an exis
ting rule-based expert system to derive a dataset consisting of 2,989 productive
 mechanistic steps and 6.14 million non-productive mechanistic steps. We then po
se identifying productive mechanistic steps as a ranking problem: rank potential
 orbital interactions such that the top ranked interactions yield the major prod
ucts. The machine learning implementation follows a two-stage approach, in which
 we first train atom level reactivity filters to prune 94.0% of non-productive r
eactions with less than a 0.1% false negative rate. Then, we train an ensemble o
f ranking models on pairs of interacting orbitals to learn a relative productivi
ty function over single mechanistic reactions in a given system. Without the use
 of explicit transformation patterns, the ensemble perfectly ranks the productiv
e mechanisms at the top 89.1% of the time, rising to 99.9% of the time when top
ranked lists with at most four non-productive reactions are considered. The fina
l system allows multi-step reaction prediction. Furthermore, it is generalizable
, making reasonable predictions over reactants and conditions which the rule-bas
ed expert system does not handle.
****

Sparse Features for PCA-Like Linear Regression
Christos Boutsidis, Petros Drineas, Malik Magdon-Ismail
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.
****

Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning
Michalis Titsias, Miguel Lázaro-Gredilla
We introduce a variational Bayesian inference algorithm which can be widely appl
ied to sparse linear models. The algorithm is based on the spike and slab prior
which, from a Bayesian perspective, is the golden standard for sparse inference.
 We apply the method to a general multi-task and multiple kernel learning model
in which a common set of Gaussian process functions is linearly combined with ta
sk-specific sparse weights, thus inducing relation between tasks. This model uni
fies several sparse linear models, such as generalized linear models, sparse fac
tor analysis and matrix factorization with missing values, so that the variation
al algorithm can be applied to all these cases. We demonstrate our approach in m
ulti-output Gaussian process regression, multi-class classification, image proce
ssing applications and collaborative filtering.
****

Neuronal Adaptation for Sampling-Based Probabilistic Inference in Perceptual Bis
tability
David Reichert, Peggy Series, Amos J. Storkey
It has been argued that perceptual multistability reflects probabilistic inferen
ce performed by the brain when sensory input is ambiguous. Alternatively, more t
raditional explanations of multistability refer to low-level mechanisms such as
neuronal adaptation. We employ a Deep Boltzmann Machine (DBM) model of cortical
processing to demonstrate that these two different approaches can be combined in
 the same framework. Based on recent developments in machine learning, we show h
ow neuronal adaptation can be understood as a mechanism that improves probabilis
tic, sampling-based inference. Using the ambiguous Necker cube image, we analyze

the perceptual switching exhibited by the model. We also examine the influence of spatial attention, and explore how binocular rivalry can be modeled with the same approach. Our work joins earlier studies in demonstrating how the principles underlying DBMs relate to cortical processing, and offers novel perspectives on the neural implementation of approximate probabilistic inference in the brain.
****

Reinforcement Learning using Kernel-Based Stochastic Factorization
Andre Barreto, Doina Precup, Joelle Pineau
Kernel-based reinforcement-learning (KBRL) is a method for learning a decision policy from a set of sample transitions which stands out for its strong theoretical guarantees. However, the size of the approximator grows with the number of transitions, which makes the approach impractical for large problems. In this paper we introduce a novel algorithm to improve the scalability of KBRL. We resort to a special decomposition of a transition matrix, called stochastic factorization, to fix the size of the approximator while at the same time incorporating all the information contained in the data. The resulting algorithm, kernel-based stochastic factorization (KBSF), is much faster but still converges to a unique solution. We derive a theoretical upper bound for the distance between the value functions computed by KBRL and KBSF. The effectiveness of our method is illustrated with computational experiments on four reinforcement-learning problems, including a difficult task in which the goal is to learn a neurostimulation policy to suppress the occurrence of seizures in epileptic rat brains. We empirically demonstrate that the proposed approach is able to compress the information contained in KBRL's model. Also, on the tasks studied, KBSF outperforms two of the most prominent reinforcement-learning algorithms, namely least-squares policy iteration and fitted Q-iteration.
****

Better Mini-Batch Algorithms via Accelerated Gradient Methods
Andrew Cotter, Ohad Shamir, Nati Srebro, Karthik Sridharan
Mini-batch algorithms have recently received significant attention as a way to speed-up stochastic convex optimization problems. In this paper, we study how such algorithms can be improved using accelerated gradient methods. We provide a novel analysis, which shows how standard gradient methods may sometimes be insufficient to obtain a significant speed-up. We propose a novel accelerated gradient algorithm, which deals with this deficiency, and enjoys a uniformly superior guarantee. We conclude our paper with experiments on real-world datasets, which validates our algorithm and substantiates our theoretical insights.
****

Generalizing from Several Related Classification Tasks to a New Unlabeled Sample
Gilles Blanchard, Gyemin Lee, Clayton Scott
We consider the problem of assigning class labels to an unlabeled test data set, given several labeled training data sets drawn from similar distributions. This problem arises in several applications where data distributions fluctuate because of biological, technical, or other sources of variation. We develop a distribution-free, kernel-based approach to the problem. This approach involves identifying an appropriate reproducing kernel Hilbert space and optimizing a regularized empirical risk over the space. We present generalization error analysis, describe universal kernels, and establish universal consistency of the proposed methodology. Experimental results on flow cytometry data are presented.
****

Energetically Optimal Action Potentials
Martin Stemmler, Biswa Sengupta, Simon Laughlin, Jeremy Niven
Most action potentials in the nervous system take on the form of strong, rapid, and brief voltage deflections known as spikes, in stark contrast to other action potentials, such as in the heart, that are characterized by broad voltage plateaus. We derive the shape of the neuronal action potential from first principles, by postulating that action potential generation is strongly constrained by the brain's need to minimize energy expenditure. For a given height of an action potential, the least energy is consumed when the underlying currents obey the bang-bang principle: the currents giving rise to the spike should be intense, yet sho

rt-lived, yielding spikes with sharp onsets and offsets. Energy optimality predicts features in the biophysics that are not per se required for producing the characteristic neuronal action potential: sodium currents should be extraordinarily powerful and inactivate with voltage; both potassium and sodium currents should have kinetics that have a bell-shaped voltage-dependence; and the cooperative action of multiple `gates' should start the flow of current.
****

Global Solution of Fully-Observed Variational Bayesian Matrix Factorization is Column-Wise Independent
Shinichi Nakajima, Masashi Sugiyama, S. Babacan
Variational Bayesian matrix factorization (VBMF) efficiently approximates the posterior distribution of factorized matrices by assuming matrix-wise independence of the two factors. A recent study on fully-observed VBMF showed that, under a stronger assumption that the two factorized matrices are column-wise independent, the global optimal solution can be analytically computed. However, it was not clear how restrictive the column-wise independence assumption is. In this paper, we prove that the global solution under matrix-wise independence is actually column-wise independent, implying that the column-wise independence assumption is harmless. A practical consequence of our theoretical finding is that the global solution under matrix-wise independence (which is a standard setup) can be obtained analytically in a computationally very efficient way without any iterative algorithms. We experimentally illustrate advantages of using our analytic solution in probabilistic principal component analysis.
****

Algorithms and hardness results for parallel large margin learning
Phil Long, Rocco Servedio
We study the fundamental problem of learning an unknown large-margin halfspace in the context of parallel computation. Our main positive result is a parallel algorithm for learning a large-margin halfspace that is based on interior point methods from convex optimization and fast parallel algorithms for matrix computations. We show that this algorithm learns an unknown gamma-margin halfspace over n dimensions using poly(n,1/gamma) processors and runs in time ~O(1/gamma) + O(log n). In contrast, naive parallel algorithms that learn a gamma-margin halfspace in time that depends polylogarithmically on n have Omega(1/gamma^2) runtime dependence on gamma. Our main negative result deals with boosting, which is a standard approach to learning large-margin halfspaces. We give an information-theoretic proof that in the original PAC framework, in which a weak learning algorithm is provided as an oracle that is called by the booster, boosting cannot be parallelized: the ability to call the weak learner multiple times in parallel within a single boosting stage does not reduce the overall number of successive stages of boosting that are required.
****

Semi-supervised Regression via Parallel Field Regularization
Binbin Lin, Chiyuan Zhang, Xiaofei He
This paper studies the problem of semi-supervised learning from the vector field perspective. Many of the existing work use the graph Laplacian to ensure the smoothness of the prediction function on the data manifold. However, beyond smoothness, it is suggested by recent theoretical work that we should ensure second order smoothness for achieving faster rates of convergence for semi-supervised regression problems. To achieve this goal, we show that the second order smoothness measures the linearity of the function, and the gradient field of a linear function has to be a parallel vector field. Consequently, we propose to find a function which minimizes the empirical error, and simultaneously requires its gradient field to be as parallel as possible. We give a continuous objective function on the manifold and discuss how to discretize it by using random points. The discretized optimization problem turns out to be a sparse linear system which can be solved very efficiently. The experimental results have demonstrated the effectiveness of our proposed approach.
****

Thinning Measurement Models and Questionnaire Design

Ricardo Silva
Inferring key unobservable features of individuals is an important task in the a pplied sciences. In particular, an important source of data in fields such as ma rketing, social sciences and medicine is questionnaires: answers in such questio nnaires are noisy measures of target unobserved features. While comprehensive su rveys help to better estimate the latent variables of interest, aiming at a high number of questions comes at a price: refusal to participate in surveys can go up, as well as the rate of missing data; quality of answers can decline; costs a ssociated with applying such questionnaires can also increase. In this paper, we cast the problem of refining existing models for questionnaire data as follows: solve a constrained optimization problem of preserving the maximum amount of in formation found in a latent variable model using only a subset of existing quest ions. The goal is to find an optimal subset of a given size. For that, we first define an information theoretical measure for quantifying the quality of a reduc ed questionnaire. Three different approximate inference methods are introduced t o solve this problem. Comparisons against a simple but powerful heuristic are pr esented.
****

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials
Philipp Krähenbühl, Vladlen Koltun
Most state-of-the-art techniques for multi-class image segmentation and labeling use conditional random ■elds de■ned over pixels or image regions. While region-level models often feature dense pairwise connectivity, pixel-level models are con- siderably larger and have only permitted sparse graph structures. In this p aper, we consider fully connected CRF models de■ned on the complete set of pixel s in an image. The resulting graphs have billions of edges, making traditional i nference algorithms impractical. Our main contribution is a highly ef■cient appr oximate inference algorithm for fully connected CRF models in which the pairwise edge potentials are de■ned by a linear combination of Gaussian kernels. Our exp eri- ments demonstrate that dense connectivity at the pixel level substantially improves segmentation and labeling accuracy.
****

Simultaneous Sampling and Multi-Structure Fitting with Adaptive Reversible Jump MCMC
Trung Pham, Tat-jun Chin, Jin Yu, David Suter
Multi-structure model fitting has traditionally taken a two-stage approach: Firs t, sample a (large) number of model hypotheses, then select the subset of hypoth eses that optimise a joint fitting and model selection criterion. This disjoint two-stage approach is arguably suboptimal and inefficient - if the random sampli ng did not retrieve a good set of hypotheses, the optimised outcome will not rep resent a good fit. To overcome this weakness we propose a new multi-structure fi tting approach based on Reversible Jump MCMC. Instrumental in raising the effect iveness of our method is an adaptive hypothesis generator, whose proposal distri bution is learned incrementally and online. We prove that this adaptive proposal satisfies the diminishing adaptation property crucial for ensuring ergodicity i n MCMC. Our method effectively conducts hypothesis sampling and optimisation sim ultaneously, and gives superior computational efficiency over other methods.
****

Active dendrites: adaptation to spike-based communication
Balazs Ujfalussy, Máté Lengyel
Computational analyses of dendritic computations often assume stationary inputs to neurons, ignoring the pulsatile nature of spike-based communication between n eurons and the moment-to-moment fluctuations caused by such spiking inputs. Conv ersely, circuit computations with spiking neurons are usually formalized without regard to the rich nonlinear nature of dendritic processing. Here we address th e computational challenge faced by neurons that compute and represent analogue q uantities but communicate with digital spikes, and show that reliable computatio n of even purely linear functions of inputs can require the interplay of strongl y nonlinear subunits within the postsynaptic dendritic tree. Our theory predicts a matching of dendritic nonlinearities and synaptic weight distributions to the

joint statistics of presynaptic inputs. This approach suggests normative roles for some puzzling forms of nonlinear dendritic dynamics and plasticity.
****

Shaping Level Sets with Submodular Functions
Francis Bach
We consider a class of sparsity-inducing regularization terms based on submodular functions. While previous work has focused on non-decreasing functions, we explore symmetric submodular functions and their \lova extensions. We show that the Lovasz extension may be seen as the convex envelope of a function that depends on level sets (i.e., the set of indices whose corresponding components of the underlying predictor are greater than a given constant): this leads to a class of convex structured regularization terms that impose prior knowledge on the level sets, and not only on the supports of the underlying predictors. We provide a unified set of optimization algorithms, such as proximal operators, and theoretical guarantees (allowed level sets and recovery conditions). By selecting specific submodular functions, we give a new interpretation to known norms, such as the total variation; we also define new norms, in particular ones that are based on order statistics with application to clustering and outlier detection, and on noisy cuts in graphs with application to change point detection in the presence of outliers.
****

Probabilistic amplitude and frequency demodulation
Richard Turner, Maneesh Sahani
A number of recent scientific and engineering problems require signals to be decomposed into a product of a slowly varying positive envelope and a quickly varying carrier whose instantaneous frequency also varies slowly over time. Although signal processing provides algorithms for so-called amplitude- and frequency-demodulation (AFD), there are well known problems with all of the existing methods. Motivated by the fact that AFD is ill-posed, we approach the problem using probabilistic inference. The new approach, called probabilistic amplitude and frequency demodulation (PAFD), models instantaneous frequency using an auto-regressive generalization of the von Mises distribution, and the envelopes using Gaussian auto-regressive dynamics with a positivity constraint. A novel form of expectation propagation is used for inference. We demonstrate that although PAFD is computationally demanding, it outperforms previous approaches on synthetic and real signals in clean, noisy and missing data settings.
****

Multi-Bandit Best Arm Identification
Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, Sébastien Bubeck
We study the problem of identifying the best arm in each of the bandits in a multi-bandit multi-armed setting. We first propose an algorithm called Gap-based Exploration (GapE) that focuses on the arms whose mean is close to the mean of the best arm in the same bandit (i.e., small gap). We then introduce an algorithm, called GapE-V, which takes into account the variance of the arms in addition to their gap. We prove an upper-bound on the probability of error for both algorithms. Since GapE and GapE-V need to tune an exploration parameter that depends on the complexity of the problem, which is often unknown in advance, we also introduce variations of these algorithms that estimate this complexity online. Finally, we evaluate the performance of these algorithms and compare them to other allocation strategies on a number of synthetic problems.
****

Nonlinear Inverse Reinforcement Learning with Gaussian Processes
Sergey Levine, Zoran Popovic, Vladlen Koltun
We present a probabilistic algorithm for nonlinear inverse reinforcement learning. The goal of inverse reinforcement learning is to learn the reward function in a Markov decision process from expert demonstrations. While most prior inverse reinforcement learning algorithms represent the reward as a linear combination of a set of features, we use Gaussian processes to learn the reward as a nonlinear function, while also determining the relevance of each feature to the expert's policy. Our probabilistic algorithm allows complex behaviors to be captured fro

m suboptimal stochastic demonstrations, while automatically balancing the simplicity of the learned reward structure against its consistency with the observed actions.
****

EigenNet: A Bayesian hybrid of generative and conditional models for sparse learning
Feng Yan, Yuan Qi
For many real-world applications, we often need to select correlated variables---such as genetic variations and imaging features associated with Alzheimer's disease---in a high dimensional space. The correlation between variables presents a challenge to classical variable selection methods. To address this challenge, the elastic net has been developed and successfully applied to many applications. Despite its great success, the elastic net does not exploit the correlation information embedded in the data to select correlated variables. To overcome this limitation, we present a novel hybrid model, EigenNet, that uses the eigenstructures of data to guide variable selection. Specifically, it integrates a sparse conditional classification model with a generative model capturing variable correlations in a principled Bayesian framework. We develop an efficient active-set algorithm to estimate the model via evidence maximization. Experiments on synthetic data and imaging genetics data demonstrated the superior predictive performance of the EigenNet over the lasso, the elastic net, and the automatic relevance determination.
****

Iterative Learning for Reliable Crowdsourcing Systems
David Karger, Sewoong Oh, Devavrat Shah
Crowdsourcing systems, in which tasks are electronically distributed to numerous ``information piece-workers'', have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting. In this paper, we consider a general model of such rowdsourcing tasks, and pose the problem of minimizing the total price (i.e., number of task assignments) that must be paid to achieve a target overall reliability. We give new algorithms for deciding which tasks to assign to which workers and for inferring correct answers from the workers’ answers. We show that our algorithm significantly outperforms majority voting and, in fact, are asymptotically optimal through comparison to an oracle that knows the reliability of every worker.
****

Minimax Localization of Structural Information in Large Noisy Matrices
Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, Aarti Singh
We consider the problem of identifying a sparse set of relevant columns and rows in a large data matrix with highly corrupted entries. This problem of identifying groups from a collection of bipartite variables such as proteins and drugs, biological species and gene sequences, malware and signatures, etc is commonly referred to as biclustering or co-clustering. Despite its great practical relevance, and although several ad-hoc methods are available for biclustering, theoretical analysis of the problem is largely non-existent. The problem we consider is also closely related to structured multiple hypothesis testing, an area of statistics that has recently witnessed a flurry of activity. We make the following contributions: i) We prove lower bounds on the minimum signal strength needed for successful recovery of a bicluster as a function of the noise variance, size of the matrix and bicluster of interest. ii) We show that a combinatorial procedure based on the scan statistic achieves this optimal limit. iii) We characterize the SNR required by several computationally tractable procedures for biclustering including element-wise thresholding, column/row average thresholding and a convex relaxation approach to sparse singular vector decomposition.
****

Crowdclustering

Ryan Gomes, Peter Welinder, Andreas Krause, Pietro Perona
Is it possible to crowdsource categorization? Amongst the challenges: (a) each a nnotator has only a partial view of the data, (b) different annotators may have different clustering criteria and may produce different numbers of categories, (c) the underlying category structure may be hierarchical. We propose a Bayesian model of how annotators may approach clustering and show how one may infer clust ers/categories, as well as annotator parameters, using this model. Our experimen ts, carried out on large collections of images, suggest that Bayesian crowdclust ering works well and may be superior to single-expert annotations.
****

Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for H MMs
Armen Allahverdyan, Aram Galstyan
We present an asymptotic analysis of Viterbi Training (VT) and contrast it with a more conventional Maximum Likelihood (ML) approach to parameter estimation in Hidden Markov Models. While ML estimator works by (locally) maximizing the likel ihood of the observed data, VT seeks to maximize the probability of the most lik ely hidden state sequence. We develop an analytical framework based on a generat ing function formalism and illustrate it on an exactly solvable model of HMM wit h one unambiguous symbol. For this particular model the ML objective function is continuously degenerate. VT objective, in contrast, is shown to have only finit e degeneracy. Furthermore, VT converges faster and results in sparser (simpler) models, thus realizing an automatic Occam's razor for HMM learning. For more gen eral scenario VT can be worse compared to ML but still capable of correctly reco vering most of the parameters.
****

Environmental statistics and the trade-off between model-based and TD learning i n humans
Dylan Simon, Nathaniel Daw
There is much evidence that humans and other animals utilize a combination of mo del-based and model-free RL methods. Although it has been proposed that these s ystems may dominate according to their relative statistical efficiency in differ ent circumstances, there is little specific evidence -- especially in humans -- as to the details of this trade-off. Accordingly, we examine the relative perfo rmance of different RL approaches under situations in which the statistics of re ward are differentially noisy and volatile. Using theory and simulation, we sho w that model-free TD learning is relatively most disadvantaged in cases of high volatility and low noise. We present data from a decision-making experiment man ipulating these parameters, showing that humans shift learning strategies in acc ord with these predictions. The statistical circumstances favoring model-based RL are also those that promote a high learning rate, which helps explain why, in psychology, the distinction between these strategies is traditionally conceived in terms of rule-based vs. incremental learning.
****

A Model for Temporal Dependencies in Event Streams
Asela Gunawardana, Christopher Meek, Puyang Xu
We introduce the Piecewise-Constant Conditional Intensity Model, a model for lea rning temporal dependencies in event streams. We describe a closed-form Bayesia n approach to learning these models, and describe an importance sampling algorit hm for forecasting future events using these models, using a proposal distributi on based on Poisson superposition. We then use synthetic data, supercomputer ev ent logs, and web search query logs to illustrate that our learning algorithm ca n efficiently learn nonlinear temporal dependencies, and that our importance sam pling algorithm can effectively forecast future events.
****

Inferring Interaction Networks using the IBP applied to microRNA Target Predicti on
Hai-son Le, Ziv Bar-joseph
Determining interactions between entities and the overall organization and clust ering of nodes in networks is a major challenge when analyzing biological and so

cial network data. Here we extend the Indian Buffet Process (IBP), a nonparametric Bayesian model, to integrate noisy interaction scores with properties of individual entities for inferring interaction networks and clustering nodes within these networks. We present an application of this method to study how microRNAs regulate mRNAs in cells. Analysis of synthetic and real data indicates that the method improves upon prior methods, correctly recovers interactions and clusters, and provides accurate biological predictions.
****

Learning unbelievable probabilities
Zachary Pitkow, Yashar Ahmadian, Ken Miller
Loopy belief propagation performs approximate inference on graphical models with loops. One might hope to compensate for the approximation by adjusting model parameters. Learning algorithms for this purpose have been explored previously, and the claim has been made that every set of locally consistent marginals can arise from belief propagation run on a graphical model. On the contrary, here we show that many probability distributions have marginals that cannot be reached by belief propagation using any set of model parameters or any learning algorithm. We call such marginals `unbelievable.' This problem occurs whenever the Hessian of the Bethe free energy is not positive-definite at the target marginals. All learning algorithms for belief propagation necessarily fail in these cases, producing beliefs or sets of beliefs that may even be worse than the pre-learning approximation. We then show that averaging inaccurate beliefs, each obtained from belief propagation using model parameters perturbed about some learned mean values, can achieve the unbelievable marginals.
****

Relative Density-Ratio Estimation for Robust Distribution Comparison
Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, Masashi Sugiyama
Divergence estimators based on direct approximation of density-ratios without going through separate approximation of numerator and denominator densities have been successfully applied to machine learning tasks that involve distribution comparison such as outlier detection, transfer learning, and two-sample homogeneity test. However, since density-ratio functions often possess high fluctuation, divergence estimation is still a challenging task in practice. In this paper, we propose to use relative divergences for distribution comparison, which involves approximation of relative density-ratios. Since relative density-ratios are always smoother than corresponding ordinary density-ratios, our proposed method is favorable in terms of the non-parametric convergence speed. Furthermore, we show that the proposed divergence estimator has asymptotic variance independent of the model complexity under a parametric setup, implying that the proposed estimator hardly overfits even with complex models. Through experiments, we demonstrate the usefulness of the proposed approach.
****

The Manifold Tangent Classifier
Salah Rifai, Yann N. Dauphin, Pascal Vincent, Yoshua Bengio, Xavier Muller
We combine three important ideas present in previous work for building classi- ■ ers: the semi-supervised hypothesis (the input distribution contains information about the classi■er), the unsupervised manifold hypothesis (data density concen - trates near low-dimensional manifolds), and the manifold hypothesis for classi ■- cation (different classes correspond to disjoint manifolds separated by low den- sity). We exploit a novel algorithm for capturing manifold structure (high-order contractive auto-encoders) and we show how it builds a topological atlas of charts, each chart being characterized by the principal singular vectors of the Jacobian of a representation mapping. This representation learning algorithm can be stacked to yield a deep architecture, and we combine it with a domain knowledge-free version of the TangentProp algorithm to encourage the classi■er to be insensitive to local directions changes along the manifold. Record-breaking classi■cation results are obtained.
****

Manifold Precis: An Annealing Technique for Diverse Sampling of Manifolds

Nitesh Shroff, Pavan Turaga, Rama Chellappa
In this paper, we consider the 'Precis' problem of sampling K representative yet diverse data points from a large dataset. This problem arises frequently in applications such as video and document summarization, exploratory data analysis, and pre-filtering. We formulate a general theory which encompasses not just traditional techniques devised for vector spaces, but also non-Euclidean manifolds, thereby enabling these techniques to shapes, human activities, textures and many other image and video based datasets. We propose intrinsic manifold measures for measuring the quality of a selection of points with respect to their representative power, and their diversity. We then propose efficient algorithms to optimize the cost function using a novel annealing-based iterative alternation algorithm. The proposed formulation is applicable to manifolds of known geometry as well as to manifolds whose geometry needs to be estimated from samples. Experimental results show the strength and generality of the proposed approach.
****

Facial Expression Transfer with Input-Output Temporal Restricted Boltzmann Machines
Matthew Zeiler, Graham W. Taylor, Leonid Sigal, Iain Matthews, Rob Fergus
We present a type of Temporal Restricted Boltzmann Machine that defines a probability distribution over an output sequence conditional on an input sequence. It shares the desirable properties of RBMs: efficient exact inference, an exponentially more expressive latent state than HMMs, and the ability to model nonlinear structure and dynamics. We apply our model to a challenging real-world graphics problem: facial expression transfer. Our results demonstrate improved performance over several baselines modeling high-dimensional 2D and 3D data.
****

Committing Bandits
Loc Bui, Ramesh Johari, Shie Mannor
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.
****

Large-Scale Category Structure Aware Image Categorization
Bin Zhao, Fei Li, Eric Xing
Most previous research on image categorization has focused on medium-scale data sets, while large-scale image categorization with millions of images from thousands of categories remains a challenge. With the emergence of structured large-scale dataset such as the ImageNet, rich information about the conceptual relationships between images, such as a tree hierarchy among various image categories, become available. As human cognition of complex visual world benefits from underlying semantic relationships between object classes, we believe a machine learning system can and should leverage such information as well for better performance. In this paper, we employ such semantic relatedness among image categories for large-scale image categorization. Specifically, a category hierarchy is utilized to properly define loss function and select common set of features for related categories. An efficient optimization method based on proximal approximation and accelerated parallel gradient method is introduced. Experimental results on a subset of ImageNet containing 1.2 million images from 1000 categories demonstrate the effectiveness and promise of our proposed approach.
****

On Causal Discovery with Cyclic Additive Noise Models
Joris M. Mooij, Dominik Janzing, Tom Heskes, Bernhard Schölkopf
We study a particular class of cyclic causal models, where each variable is a (possibly nonlinear) function of its parents and additive noise. We prove that the causal graph of such models is generically identifiable in the bivariate, Gaussian-noise case. We also propose a method to learn such models from observational data. In the acyclic case, the method reduces to ordinary regression, but in the more challenging cyclic case, an additional term arises in the loss function, which makes it a special case of nonlinear independent component analysis. We il

lustrate the proposed method on synthetic data.
****

Sparse recovery by thresholded non-negative least squares
Martin Slawski, Matthias Hein
Non-negative data are commonly encountered in numerous fields, making non-negative least squares regression (NNLS) a frequently used tool. At least relative to its simplicity, it often performs rather well in practice. Serious doubts about its usefulness arise for modern high-dimensional linear models. Even in this setting - unlike first intuition may suggest - we show that for a broad class of designs, NNLS is resistant to overfitting and works excellently for sparse recovery when combined with thresholding, experimentally even outperforming L1-regularization. Since NNLS also circumvents the delicate choice of a regularization parameter, our findings suggest that NNLS may be the method of choice.
****

A Two-Stage Weighting Framework for Multi-Source Domain Adaptation
Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, Jieping Ye
Discriminative learning when training and test data belong to different distributions is a challenging and complex task. Often times we have very few or no labeled data from the test or target distribution but may have plenty of labeled data from multiple related sources with different distributions. The difference in distributions may be in both marginal and conditional probabilities. Most of the existing domain adaptation work focuses on the marginal probability distribution difference between the domains, assuming that the conditional probabilities are similar. However in many real world applications, conditional probability distribution differences are as commonplace as marginal probability differences. In this paper we propose a two-stage domain adaptation methodology which combines weighted data from multiple sources based on marginal probability differences (first stage) as well as conditional probability differences (second stage), with the target domain data. The weights for minimizing the marginal probability differences are estimated independently, while the weights for minimizing conditional probability differences are computed simultaneously by exploiting the potential interaction among multiple sources. We also provide a theoretical analysis on the generalization performance of the proposed multi-source domain adaptation formulation using the weighted Rademacher complexity measure. Empirical comparisons with existing state-of-the-art domain adaptation methods using three real-world datasets demonstrate the effectiveness of the proposed approach.
****

Variance Reduction in Monte-Carlo Tree Search
Joel Veness, Marc Lanctot, Michael Bowling
Monte-Carlo Tree Search (MCTS) has proven to be a powerful, generic planning technique for decision-making in single-agent and adversarial environments. The stochastic nature of the Monte-Carlo simulations introduces errors in the value estimates, both in terms of bias and variance. Whilst reducing bias (typically through the addition of domain knowledge) has been studied in the MCTS literature, comparatively little effort has focused on reducing variance. This is somewhat surprising, since variance reduction techniques are a well-studied area in classical statistics. In this paper, we examine the application of some standard techniques for variance reduction in MCTS, including common random numbers, antithetic variates and control variates. We demonstrate how these techniques can be applied to MCTS and explore their efficacy on three different stochastic, single-agent settings: Pig, Can't Stop and Dominion.
****

Advice Refinement in Knowledge-Based SVMs
Gautam Kunapuli, Richard Maclin, Jude Shavlik
Knowledge-based support vector machines (KBSVMs) incorporate advice from domain experts, which can improve generalization significantly. A major limitation that has not been fully addressed occurs when the expert advice is imperfect, which can lead to poorer models. We propose a model that extends KBSVMs and is able to not only learn from data and advice, but also simultaneously improve the advice. The proposed approach is particularly effective for knowledge discovery in dom

ains with few labeled examples. The proposed model contains bilinear constraints , and is solved using two iterative approaches: successive linear programming an d a constrained concave-convex approach. Experimental results demonstrate that t hese algorithms yield useful refinements to expert advice, as well as improve th e performance of the learning algorithm overall.
****

On fast approximate submodular minimization
Stefanie Jegelka, Hui Lin, Jeff A. Bilmes
We are motivated by an application to extract a representative subset of machine learning training data and by the poor empirical performance we observe of the popular minimum norm algorithm. In fact, for our application, minimum norm can h ave a running time of about O(n^7) (O(n^5) oracle calls). We therefore propose a fast approximate method to minimize arbitrary submodular functions. For a lar ge sub-class of submodular functions, the algorithm is exact. Other submodular f unctions are iteratively approximated by tight submodular upper bounds, and then repeatedly optimized. We show theoretical properties, and empirical results sug gest significant speedups over minimum norm while retaining higher accuracies.
****

Multiple Instance Filtering
Kamil Wnuk, Stefano Soatto
We propose a robust filtering approach based on semi-supervised and multiple ins tance learning (MIL). We assume that the posterior density would be unimodal if not for the effect of outliers that we do not wish to explicitly model. Therefor e, we seek for a point estimate at the outset, rather than a generic approximati on of the entire posterior. Our approach can be thought of as a combination of s tandard finite-dimensional filtering (Extended Kalman Filter, or Unscented Filte r) with multiple instance learning, whereby the initial condition comes with a p utative set of inlier measurements. We show how both the state (regression) and the inlier set (classification) can be estimated iteratively and causally by pro cessing only the current measurement. We illustrate our approach on visual track ing problems whereby the object of interest (target) moves and evolves as a resu lt of occlusions and deformations, and partial knowledge of the target is given in the form of a bounding box (training set).
****

A reinterpretation of the policy oscillation phenomenon in approximate policy it eration
Paul Wagner
A majority of approximate dynamic programming approaches to the reinforcement le arning problem can be categorized into greedy value function methods and value-b ased policy gradient methods. The former approach, although fast, is well known to be susceptible to the policy oscillation phenomenon. We take a fresh view to this phenomenon by casting a considerable subset of the former approach as a lim iting special case of the latter. We explain the phenomenon in terms of this vie w and illustrate the underlying mechanism with artificial examples. We also use it to derive the constrained natural actor-critic algorithm that can interpolate between the aforementioned approaches. In addition, it has been suggested in th e literature that the oscillation phenomenon might be subtly connected to the gr ossly suboptimal performance in the Tetris benchmark problem of all attempted ap proximate dynamic programming methods. We report empirical evidence against such a connection and in favor of an alternative explanation. Finally, we report sco res in the Tetris problem that improve on existing dynamic programming based res ults.
****

θ-MRF: Capturing Spatial and Semantic Structure in the Parameters for Sce ne Understanding
Congcong Li, Ashutosh Saxena, Tsuhan Chen
For most scene understanding tasks (such as object detection or depth estimation), the classifiers need to consider contextual information in addition to the lo cal features. We can capture such contextual information by taking as input the features/attributes from all the regions in the image. However, this contextual

dependence also varies with the spatial location of the region of interest, and we therefore need a different set of parameters for each spatial location. This results in a very large number of parameters. In this work, we model the independence properties between the parameters for each location and for each task, by defining a Markov Random Field (MRF) over the parameters. In particular, two sets of parameters are encouraged to have similar values if they are spatially close or semantically close. Our method is, in principle, complementary to other ways of capturing context such as the ones that use a graphical model over the labels instead. In extensive evaluation over two different settings, of multi-class object detection and of multiple scene understanding tasks (scene categorization, depth estimation, geometric labeling), our method beats the state-of-the-art methods in all the four tasks.

****

Maximum Covariance Unfolding : Manifold Learning for Bimodal Data

Vijay Mahadevan, Chi Wong, Jose Pereira, Tom Liu, Nuno Vasconcelos, Lawrence Saul

We propose maximum covariance unfolding (MCU), a manifold learning algorithm for simultaneous dimensionality reduction of data from different input modalities. Given high dimensional inputs from two different but naturally aligned sources, MCU computes a common low dimensional embedding that maximizes the cross-modal (inter-source) correlations while preserving the local (intra-source) distances. In this paper, we explore two applications of MCU. First we use MCU to analyze EEG-fMRI data, where an important goal is to visualize the fMRI voxels that are most strongly correlated with changes in EEG traces. To perform this visualization, we augment MCU with an additional step for metric learning in the high dimensional voxel space. Second, we use MCU to perform cross-modal retrieval of matched image and text samples from Wikipedia. To manage large applications of MCU, we develop a fast implementation based on ideas from spectral graph theory. These ideas transform the original problem for MCU, one of semidefinite programming, into a simpler problem in semidefinite quadratic linear programming.

****

Noise Thresholds for Spectral Clustering

Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, Aarti Singh

Although spectral clustering has enjoyed considerable empirical success in machine learning, its theoretical properties are not yet fully developed. We analyze the performance of a spectral algorithm for hierarchical clustering and show that on a class of hierarchically structured similarity matrices, this algorithm can tolerate noise that grows with the number of data points while still perfectly recovering the hierarchical clusters with high probability. We additionally improve upon previous results for k-way spectral clustering to derive conditions under which spectral clustering makes no mistakes. Further, using minimax analysis, we derive tight upper and lower bounds for the clustering problem and compare the performance of spectral clustering to these information theoretic limits. We also present experiments on simulated and real world data illustrating our results.

****

Kernel Embeddings of Latent Tree Graphical Models

Le Song, Eric Xing, Ankur Parikh

Latent tree graphical models are natural tools for expressing long range and hierarchical dependencies among many variables which are common in computer vision, bioinformatics and natural language processing problems. However, existing models are largely restricted to discrete and Gaussian variables due to computational constraints; furthermore, algorithms for estimating the latent tree structure and learning the model parameters are largely restricted to heuristic local search. We present a method based on kernel embeddings of distributions for latent tree graphical models with continuous and non-Gaussian variables. Our method can recover the latent tree structures with provable guarantees and perform local-minimum free parameter learning and efficient inference. Experiments on simulated and real data show the advantage of our proposed approach.

****

Learning a Distance Metric from a Network
Blake Shaw, Bert Huang, Tony Jebara
Many real-world networks are described by both connectivity information and features for every node. To better model and understand these networks, we present structure preserving metric learning (SPML), an algorithm for learning a Mahalanobis distance metric from a network such that the learned distances are tied to the inherent connectivity structure of the network. Like the graph embedding algorithm structure preserving embedding, SPML learns a metric which is structure preserving, meaning a connectivity algorithm such as k-nearest neighbors will yield the correct connectivity when applied using the distances from the learned metric. We show a variety of synthetic and real-world experiments where SPML predicts link patterns from node features more accurately than standard techniques. We further demonstrate a method for optimizing SPML based on stochastic gradient descent which removes the running-time dependency on the size of the network and allows the method to easily scale to networks of thousands of nodes and millions of edges.
****

Selective Prediction of Financial Trends with Hidden Markov Models
Dmitry Pidan, Ran El-Yaniv
Focusing on short term trend prediction in a financial context, we consider the problem of selective prediction whereby the predictor can abstain from prediction in order to improve performance. We examine two types of selective mechanisms for HMM predictors. The first is a rejection in the spirit of Chow’s well-known ambiguity principle. The second is a specialized mechanism for HMMs that identifies low quality HMM states and abstain from prediction in those states. We call this model selective HMM (sHMM). In both approaches we can trade-off prediction coverage to gain better accuracy in a controlled manner. We compare performance of the ambiguity-based rejection technique with that of the sHMM approach. Our results indicate that both methods are effective, and that the sHMM model is superior.
****

Blending Autonomous Exploration and Apprenticeship Learning
Thomas Walsh, Daniel Hewlett, Clayton Morrison
We present theoretical and empirical results for a framework that combines the benefits of apprenticeship and autonomous reinforcement learning. Our approach modifies an existing apprenticeship learning framework that relies on teacher demonstrations and does not necessarily explore the environment. The first change is replacing previously used Mistake Bound model learners with a recently proposed framework that melds the KWIK and Mistake Bound supervised learning protocols. The second change is introducing a communication of expected utility from the student to the teacher. The resulting system only uses teacher traces when the agent needs to learn concepts it cannot efficiently learn on its own.
****

Transfer Learning by Borrowing Examples for Multiclass Object Detection
Joseph J. Lim, Russ R. Salakhutdinov, Antonio Torralba
Despite the recent trend of increasingly large datasets for object detection, there still exist many classes with few training examples. To overcome this lack of train- ing data for certain classes, we propose a novel way of augmenting the training data for each class by borrowing and transforming examples from other classes. Our model learns which training instances from other classes to borrow and how to transform the borrowed examples so that they become more similar to instances from the target class. Our experimental results demonstrate that our new object detector, with borrowed and transformed examples, improves upon the current state-of-the-art detector on the challenging SUN09 object detection dataset.
****

A blind sparse deconvolution method for neural spike identification
Chaitanya Ekanadham, Daniel Tranchina, Eero Simoncelli
We consider the problem of estimating neural spikes from extracellular voltage recordings. Most current methods are based on clustering, which requires substantial human supervision and produces systematic errors by failing to properly hand

le temporally overlapping spikes. We formulate the problem as one of statistical inference, in which the recorded voltage is a noisy sum of the spike trains of each neuron convolved with its associated spike waveform. Joint maximum-a-posteriori (MAP) estimation of the waveforms and spikes is then a blind deconvolution problem in which the coefficients are sparse. We develop a block-coordinate descent method for approximating the MAP solution. We validate our method on data simulated according to the generative model, as well as on real data for which ground truth is available via simultaneous intracellular recordings. In both cases, our method substantially reduces the number of missed spikes and false positives when compared to a standard clustering algorithm, primarily by recovering temporally overlapping spikes. The method offers a fully automated alternative to clustering methods that is less susceptible to systematic errors.

****

Clustering via Dirichlet Process Mixture Models for Portable Skill Discovery

Scott Niekum, Andrew Barto

Skill discovery algorithms in reinforcement learning typically identify single states or regions in state space that correspond to task-specific subgoals. However, such methods do not directly address the question of how many distinct skills are appropriate for solving the tasks that the agent faces. This can be highly inefficient when many identified subgoals correspond to the same underlying skill, but are all used individually as skill goals. Furthermore, skills created in this manner are often only transferable to tasks that share identical state spaces, since corresponding subgoals across tasks are not merged into a single skill goal. We show that these problems can be overcome by clustering subgoal data defined in an agent-space and using the resulting clusters as templates for skill termination conditions. Clustering via a Dirichlet process mixture model is used to discover a minimal, sufficient collection of portable skills.

****

Unsupervised learning models of primary cortical receptive fields and receptive field plasticity

Maneesh Bhand, Ritvik Mudur, Bipin Suresh, Andrew Saxe, Andrew Ng

The efficient coding hypothesis holds that neural receptive fields are adapted to the statistics of the environment, but is agnostic to the timescale of this adaptation, which occurs on both evolutionary and developmental timescales. In this work we focus on that component of adaptation which occurs during an organism's lifetime, and show that a number of unsupervised feature learning algorithms can account for features of normal receptive field properties across multiple primary sensory cortices. Furthermore, we show that the same algorithms account for altered receptive field properties in response to experimentally altered environmental statistics. Based on these modeling results we propose these models as phenomenological models of receptive field plasticity during an organism's lifetime. Finally, due to the success of the same models in multiple sensory areas, we suggest that these algorithms may provide a constructive realization of the theory, first proposed by Mountcastle (1978), that a qualitatively similar learning algorithm acts throughout primary sensory cortices.

****

Improved Algorithms for Linear Stochastic Bandits

Yasin Abbasi-yadkori, Dávid Pál, Csaba Szepesvári

We improve the theoretical analysis and empirical performance of algorithms for the stochastic multi-armed bandit problem and the linear stochastic multi-armed bandit problem. In particular, we show that a simple modification of Auer’s UCB algorithm (Auer, 2002) achieves with high probability constant regret. More importantly, we modify and, consequently, improve the analysis of the algorithm for the for linear stochastic bandit problem studied by Auer (2002), Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Li et al. (2010). Our modification improves the regret bound by a logarithmic factor, though experiments show a vast improvement. In both cases, the improvement stems from the construction of smaller confidence sets. For their construction we use a novel tail inequality for vector-valued martingales.

****

From Bandits to Experts: On the Value of Side-Observations
Shie Mannor, Ohad Shamir
We consider an adversarial online learning setting where a decision maker can choose an action in every stage of the game. In addition to observing the reward of the chosen action, the decision maker gets side observations on the reward he would have obtained had he chosen some of the other actions. The observation structure is encoded as a graph, where node i is linked to node j if sampling i provides information on the reward of j. This setting naturally interpolates between the well-known ``experts'' setting, where the decision maker can view all rewards, and the multi-armed bandits setting, where the decision maker can only view the reward of the chosen action. We develop practical algorithms with provable regret guarantees, which depend on non-trivial graph-theoretic properties of the information feedback structure. We also provide partially-matching lower bounds.
****
Optimal Reinforcement Learning for Gaussian Systems
Philipp Hennig
The exploration-exploitation trade-off is among the central challenges of reinforcement learning. The optimal Bayesian solution is intractable in general. This paper studies to what extent analytic statements about optimal learning are possible if all beliefs are Gaussian processes. A first order approximation of learning of both loss and dynamics, for nonlinear, time-varying systems in continuous time and space, subject to a relatively weak restriction on the dynamics, is described by an infinite-dimensional partial differential equation. An approximate finite-dimensional projection gives an impression for how this result may be helpful.
****
An Empirical Evaluation of Thompson Sampling
Olivier Chapelle, Lihong Li
Thompson sampling is one of oldest heuristic to address the exploration / exploitation trade-off, but it is surprisingly not very popular in the literature. We present here some empirical results using Thompson sampling on simulated and real data, and show that it is highly competitive. And since this heuristic is very easy to implement, we argue that it should be part of the standard baselines to compare against.
****
Efficient Offline Communication Policies for Factored Multiagent POMDPs
João Messias, Matthijs Spaan, Pedro Lima
Factored Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) form a powerful framework for multiagent planning under uncertainty, but optimal solutions require a rigid history-based policy representation. In this paper we allow inter-agent communication which turns the problem in a centralized Multiagent POMDP (MPOMDP). We map belief distributions over state factors to an agent's local actions by exploiting structure in the joint MPOMDP policy. The key point is that when sparse dependencies between the agents' decisions exist, often the belief over its local state factors is sufficient for an agent to unequivocally identify the optimal action, and communication can be avoided. We formalize these notions by casting the problem into convex optimization form, and present experimental results illustrating the savings in communication that we can obtain.
****
Maximal Cliques that Satisfy Hard Constraints with Application to Deformable Object Model Learning
Xinggang Wang, Xiang Bai, Xingwei Yang, Wenyu Liu, Longin Latecki
We propose a novel inference framework for finding maximal cliques in a weighted graph that satisfy hard constraints. The constraints specify the graph nodes that must belong to the solution as well as mutual exclusions of graph nodes, i.e., sets of nodes that cannot belong to the same solution. The proposed inference is based on a novel particle filter algorithm with state permeations. We apply the inference framework to a challenging problem of learning part-based, deform

able object models. Two core problems in the learning framework, matching of ima ge patches and finding salient parts, are formulated as two instances of the pro blem of finding maximal cliques with hard constraints. Our learning framework yi elds discriminative part based object models that achieve very good detection ra te, and outperform other methods on object classes with large deformation.
****

Quasi-Newton Methods for Markov Chain Monte Carlo
Yichuan Zhang, Charles Sutton
The performance of Markov chain Monte Carlo methods is often sensitive to the sc aling and correlations between the random variables of interest. An important so urce of information about the local correlation and scale is given by the Hessia n matrix of the target distribution, but this is often either computationally ex pensive or infeasible. In this paper we propose MCMC samplers that make use of q uasi-Newton approximations from the optimization literature, that approximate th e Hessian of the target distribution from previous samples and gradients generat ed by the sampler. A key issue is that MCMC samplers that depend on the history of previous states are in general not valid. We address this problem by using li mited memory quasi-Newton methods, which depend only on a fixed window of previo us samples. On several real world datasets, we show that the quasi-Newton sample r is a more effective sampler than standard Hamiltonian Monte Carlo at a fractio n of the cost of MCMC methods that require higher-order derivatives.
****

Inference in continuous-time change-point models
Florian Stimberg, Manfred Opper, Guido Sanguinetti, Andreas Ruttor
We consider the problem of Bayesian inference for continuous time multi-stable s tochastic systems which can change both their diffusion and drift parameters at discrete times. We propose exact inference and sampling methodologies for two sp ecific cases where the discontinuous dynamics is given by a Poisson process and a two-state Markovian switch. We test the methodology on simulated data, and app ly it to two real data sets in finance and systems biology. Our experimental res ults show that the approach leads to valid inferences and non-trivial insights.
****

Universal low-rank matrix recovery from Pauli measurements
Yi-kai Liu
We study the problem of reconstructing an unknown matrix M of rank r and dimensi on d using O(rd polylog d) Pauli measurements. This has applications in quantum state tomography, and is a non-commutative analogue of a well-known problem in compressed sensing: recovering a sparse vector from a few of its Fourier coeffi cients. We show that almost all sets of O(rd log^6 d) Pauli measurements sa tisfy the rank-r restricted isometry property (RIP). This implies that M can be recovered from a fixed ("universal") set of Pauli measurements, using nuclear-n orm minimization (e.g., the matrix Lasso), with nearly-optimal bounds on the err or. A similar result holds for any class of measurements that use an orthonorma l operator basis whose elements have small operator norm. Our proof uses Dudley 's inequality for Gaussian processes, together with bounds on covering numbers o btained via entropy duality.
****

A Convergence Analysis of Log-Linear Training
Simon Wiesler, Hermann Ney
Log-linear models are widely used probability models for statistical pattern rec ognition. Typically, log-linear models are trained according to a convex criteri on. In recent years, the interest in log-linear models has greatly increased. Th e optimization of log-linear model parameters is costly and therefore an importa nt topic, in particular for large-scale applications. Different optimization alg orithms have been evaluated empirically in many papers. In this work, we analyze the optimization problem analytically and show that the training of log-linear models can be highly ill-conditioned. We verify our findings on two handwriting tasks. By making use of our convergence analysis, we obtain good results on a la rge-scale continuous handwriting recognition task with a simple and generic appr oach.

****
On the accuracy of l1-filtering of signals with block-sparse structure
Fatma Karzan, Arkadi S. Nemirovski, Boris Polyak, Anatoli Juditsky
We discuss new methods for the recovery of signals with block-sparse structure, based on l1-minimization. Our emphasis is on the efficiently computable error bo unds for the recovery routines. We optimize these bounds with respect to the met hod parameters to construct the estimators with improved statistical properties. We justify the proposed approach with an oracle inequality which links the prop erties of the recovery algorithms and the best estimation performance.
****
Group Anomaly Detection using Flexible Genre Models
Liang Xiong, Barnabás Póczos, Jeff Schneider
An important task in exploring and analyzing real-world data sets is to detect u nusual and interesting phenomena. In this paper, we study the group anomaly dete ction problem. Unlike traditional anomaly detection research that focuses on dat a points, our goal is to discover anomalous aggregated behaviors of groups of po ints. For this purpose, we propose the Flexible Genre Model (FGM). FGM is design ed to characterize data groups at both the point level and the group level so as to detect various types of group anomalies. We evaluate the effectiveness of FG M on both synthetic and real data sets including images and turbulence data, and show that it is superior to existing approaches in detecting group anomalies.
****
Randomized Algorithms for Comparison-based Search
Dominique Tschopp, Suhas Diggavi, Payam Delgosha, Soheil Mohajer
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.
****
Multiple Instance Learning on Structured Data
Dan Zhang, Yan Liu, Luo Si, Jian Zhang, Richard Lawrence
Most existing Multiple-Instance Learning (MIL) algorithms assume data instances and/or data bags are independently and identically distributed. But there ofte n exists rich additional dependency/structure information between instances/bag s within many applications of MIL. Ignoring this structure information limits t he performance of existing MIL algorithms. This paper explores the research pr oblem as multiple instance learning on structured data (MILSD) and formulates a novel framework that considers additional structure information. In particular , an effective and efficient optimization algorithm has been proposed to solve the original non-convex optimization problem by using a combination of Concave -Convex Constraint Programming (CCCP) method and an adapted Cutting Plane metho d, which deals with two sets of constraints caused by learning on instances wi thin individual bags and learning on structured data. Our method has the nice c onvergence property, with specified precision on each set of constraints. Exper imental results on three different applications, i.e., webpage classification, market targeting, and protein fold identification, clearly demonstrate the adv antages of the proposed method over state-of-the-art methods.
****
Probabilistic Joint Image Segmentation and Labeling
Adrian Ion, Joao Carreira, Cristian Sminchisescu
We present a joint image segmentation and labeling model (JSL) which, given a ba g of figure-ground segment hypotheses extracted at multiple image locations and scales, constructs a joint probability distribution over both the compatible ima ge interpretations (tilings or image segmentations) composed from those segments , and over their labeling into categories. The process of drawing samples from the joint distribution can be interpreted as first sampling tilings, modeled as maximal cliques, from a graph connecting spatially non-overlapping segments in t he bag, followed by sampling labels for those segments, conditioned on the choic e of a particular tiling. We learn the segmentation and labeling parameters join tly, based on Maximum Likelihood with a novel Incremental Saddle Point estimatio

n procedure. The partition function over tilings and labelings is increasingly m
ore accurately approximated by including incorrect configurations that a not-yet
-competent model rates probable during learning. We show that the proposed met
hodology matches the current state of the art in the Stanford dataset, as well a
s in VOC2010, where 41.7% accuracy on the test set is achieved.
****

Learning to Search Efficiently in High Dimensions

Zhen Li, Huazhong Ning, Liangliang Cao, Tong Zhang, Yihong Gong, Thomas S. Huang

High dimensional similarity search in large scale databases becomes an important
 challenge due to the advent of Internet. For such applications, specialized dat
a structures are required to achieve computational efficiency. Traditional appro
aches relied on algorithmic constructions that are often data independent (such
as Locality Sensitive Hashing) or weakly dependent (such as kd-trees, k-means tr
ees). While supervised learning algorithms have been applied to related problems
, those proposed in the literature mainly focused on learning hash codes optimiz
ed for compact embedding of the data rather than search efficiency. Consequently
 such an embedding has to be used with linear scan or another search algorithm.
Hence learning to hash does not directly address the search efficiency issue. Th
is paper considers a new framework that applies supervised learning to directly
optimize a data structure that supports efficient large scale search. Our approa
ch takes both search quality and computational cost into consideration. Specific
ally, we learn a boosted search forest that is optimized using pair-wise similar
ity labeled examples. The output of this search forest can be efficiently conver
ted into an inverted indexing data structure, which can leverage modern text sea
rch infrastructure to achieve both scalability and efficiency. Experimental resu
lts show that our approach significantly outperforms the start-of-the-art learni
ng to hash methods (such as spectral hashing), as well as state-of-the-art high
dimensional search algorithms (such as LSH and k-means trees).
****

Learning a Tree of Metrics with Disjoint Visual Features

Kristen Grauman, Fei Sha, Sung Hwang

We introduce an approach to learn discriminative visual representations while ex
ploiting external semantic knowledge about object category relationships. Given
 a hierarchical taxonomy that captures semantic similarity between the objects,
we learn a corresponding tree of metrics (ToM). In this tree, we have one metri
c for each non-leaf node of the object hierarchy, and each metric is responsible
 for discriminating among its immediate subcategory children. Specifically, a M
ahalanobis metric learned for a given node must satisfy the appropriate (dis)sim
ilarity constraints generated only among its subtree members' training instances
. To further exploit the semantics, we introduce a novel regularizer coupling t
he metrics that prefers a sparse disjoint set of features to be selected for eac
h metric relative to its ancestor supercategory nodes' metrics. Intuitively, th
is reflects that visual cues most useful to distinguish the generic classes (e.g
., feline vs. canine) should be different than those cues most useful to disting
uish their component fine-grained classes (e.g., Persian cat vs. Siamese cat).
We validate our approach with multiple image datasets using the WordNet taxonomy
, show its advantages over alternative metric learning approaches, and analyze t
he meaning of attribute features selected by our algorithm.
****

Monte Carlo Value Iteration with Macro-Actions

Zhan Lim, Lee Sun, David Hsu

POMDP planning faces two major computational challenges: large state spaces and
long planning horizons. The recently introduced Monte Carlo Value Iteration (MCV
I) can tackle POMDPs with very large discrete state spaces or continuous state s
paces, but its performance degrades when faced with long planning horizons. This
 paper presents Macro-MCVI, which extends MCVI by exploiting macro-actions for t
emporal abstraction. We provide sufficient conditions for Macro-MCVI to inherit
the good theoretical properties of MCVI. Macro-MCVI does not require explicit co
nstruction of probabilistic models for macro-actions and is thus easy to apply i
n practice. Experiments show that Macro-MCVI substantially improves the performa

nce of MCVI with suitable macro-actions.
****
Non-parametric Group Orthogonal Matching Pursuit for Sparse Learning with Multip
le Kernels
Vikas Sindhwani, Aurelie C. Lozano
We consider regularized risk minimization in a large dictionary of Reproducing
kernel Hilbert Spaces (RKHSs) over which the target function has a sparse repres
entation. This setting, commonly referred to as Sparse Multiple Kernel Learning
(MKL), may be viewed as the non-parametric extension of group sparsity in line
ar models. While the two dominant algorithmic strands of sparse learning, namel
y convex relaxations using l1 norm (e.g., Lasso) and greedy methods (e.g., OMP)
, have both been rigorously extended for group sparsity, the sparse MKL literat
ure has so farmainly adopted the former withmild empirical success. In this pap
er, we close this gap by proposing a Group-OMP based framework for sparse multi
ple kernel learning. Unlike l1-MKL, our approach decouples the sparsity regular
izer (via a direct l0 constraint) from the smoothness regularizer (via RKHS nor
ms) which leads to better empirical performance as well as a simpler optimizati
on procedure that only requires a black-box single-kernel solver. The algorithm
ic development and empirical studies are complemented by theoretical analyses i
n terms of Rademacher generalization bounds and sparse recovery conditions anal
ogous to those for OMP [27] and Group-OMP [16].
****
Distributed Delayed Stochastic Optimization
Alekh Agarwal, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
An Unsupervised Decontamination Procedure For Improving The Reliability Of Human
Judgments
Michael C. Mozer, Benjamin Link, Harold Pashler
Psychologists have long been struck by individuals' limitations in expressing th
eir internal sensations, impressions, and evaluations via rating scales. Instea
d of using an absolute scale, individuals rely on reference points from recent e
xperience. This relativity of judgment limits the informativeness of responses
on surveys, questionnaires, and evaluation forms. Fortunately, the cognitive pr
ocesses that map stimuli to responses are not simply noisy, but rather are influ
enced by recent experience in a lawful manner. We explore techniques to remove
sequential dependencies, and thereby decontaminate a series of ratings to obtai
n more meaningful human judgments. In our formulation, the problem is to infer
latent (subjective) impressions from a sequence of stimulus labels (e.g., movie
names) and responses. We describe an unsupervised approach that simultaneously
recovers the impressions and parameters of a contamination model that predicts
how recent judgments affect the current response. We test our iterated impre
ssion inference, or I^3, algorithm in three domains: rating the gap between dot
s, the desirability of a movie based on an advertisement, and the morality of an
action. We demonstrate significant objective improvements in the quality of th
e recovered impressions.
****
Contextual Gaussian Process Bandit Optimization
Andreas Krause, Cheng Ong
How should we design experiments to maximize performance of a complex system, ta
king into account uncontrollable environmental conditions? How should we select
relevant documents (ads) to display, given information about the user? These tas
ks can be formalized as contextual bandit problems, where at each round, we rece
ive context (about the experimental conditions, the query), and have to choose a
n action (parameters, documents). The key challenge is to trade off exploration
by gathering data for estimating the mean payoff function over the context-actio
n space, and to exploit by choosing an action deemed optimal based on the gather

ed data. We model the payoff function as a sample from a Gaussian process define d over the joint context-action space, and develop CGP-UCB, an intuitive upper-c onfidence style algorithm. We show that by mixing and matching kernels for conte xts and actions, CGP-UCB can handle a variety of practical applications. We furt her provide generic tools for deriving regret bounds when using such composite k ernel functions. Lastly, we evaluate our algorithm on two case studies, in the c ontext of automated vaccine design and sensor management. We show that context-s ensitive optimization outperforms no or naive use of context.
****

Learning with the weighted trace-norm under arbitrary sampling distributions
Rina Foygel, Ohad Shamir, Nati Srebro, Russ R. Salakhutdinov
We provide rigorous guarantees on learning with the weighted trace-norm under ar bitrary sampling distributions. We show that the standard weighted-trace norm m ight fail when the sampling distribution is not a product distribution (i.e. whe n row and column indexes are not selected independently), present a corrected va riant for which we establish strong learning guarantees, and demonstrate that it works better in practice. We provide guarantees when weighting by either the t rue or empirical sampling distribution, and suggest that even if the true distri bution is known (or is uniform), weighting by the empirical distribution may be beneficial.
****

Demixed Principal Component Analysis
Wieland Brendel, Ranulfo Romo, Christian K. Machens
In many experiments, the data points collected live in high-dimensional observat ion spaces, yet can be assigned a set of labels or parameters. In electrophysiol ogical recordings, for instance, the responses of populations of neurons general ly depend on mixtures of experimentally controlled parameters. The heterogeneity and diversity of these parameter dependencies can make visualization and interp retation of such data extremely difficult. Standard dimensionality reduction tec hniques such as principal component analysis (PCA) can provide a succinct and co mplete description of the data, but the description is constructed independent o f the relevant task variables and is often hard to interpret. Here, we start wit h the assumption that a particularly informative description is one that reveals the dependency of the high-dimensional data on the individual parameters. We sh ow how to modify the loss function of PCA so that the principal components seek to capture both the maximum amount of variance about the data, while also depend ing on a minimum number of parameters. We call this method demixed principal com ponent analysis (dPCA) as the principal components here segregate the parameter dependencies. We phrase the problem as a probabilistic graphical model, and pres ent a fast Expectation-Maximization (EM) algorithm. We demonstrate the use of th is algorithm for electrophysiological data and show that it serves to demix the parameter-dependence of a neural population response.
****

Reconstructing Patterns of Information Diffusion from Incomplete Observations
Flavio Chierichetti, David Liben-nowell, Jon Kleinberg
Motivated by the spread of on-line information in general and on-line petition s in particular, recent research has raised the following combinatorial estima tion problem. There is a tree T that we cannot observe directly (representing the structure along which the information has spread), and certain nodes random ly decide to make their copy of the information public. In the case of a petiti on, the list of names on each public copy of the petition also reveals a path leading back to the root of the tree. What can we conclude about the properties of the tree we observe from these revealed paths, and can we use the structur e of the observed tree to estimate the size of the full unobserved tree T? H ere we provide the first algorithm for this size estimation task, together with provable guarantees on its performance. We also establish structural propertie s of the observed tree, providing the first rigorous explanation for some of th e unusual structural phenomena present in the spread of real chain-letter petit ions on the Internet.
****

Differentially Private M-Estimators
Jing Lei
This paper studies privacy preserving M-estimators using perturbed histograms. T he proposed approach allows the release of a wide class of M-estimators with bo th differential privacy and statistical utility without knowing a priori the par ticular inference procedure. The performance of the proposed method is demonstra ted through a careful study of the convergence rates. A practical algorithm is g iven and applied on a real world data set containing both continuous and categor ical variables.
****
Target Neighbor Consistent Feature Weighting for Nearest Neighbor Classification
Ichiro Takeuchi, Masashi Sugiyama
We consider feature selection and weighting for nearest neighbor classifiers. A technical challenge in this scenario is how to cope with the discrete update of nearest neighbors when the feature space metric is changed during the learning p rocess. This issue, called the target neighbor change, was not properly addresse d in the existing feature weighting and metric learning literature. In this pap er, we propose a novel feature weighting algorithm that can exactly and efficien tly keep track of the correct target neighbors via sequential quadratic programm ing. To the best of our knowledge, this is the first algorithm that guarantees t he consistency between target neighbors and the feature space metric. We further show that the proposed algorithm can be naturally combined with regularization path tracking, allowing computationally efficient selection of the regularizatio n parameter. We demonstrate the effectiveness of the proposed algorithm through experiments.
****
Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Al gorithms
Liefeng Bo, Xiaofeng Ren, Dieter Fox
Extracting good representations from images is essential for many computer visio n tasks. In this paper, we propose hierarchical matching pursuit (HMP), which bu ilds a feature hierarchy layer-by-layer using an efficient matching pursuit enco der. It includes three modules: batch (tree) orthogonal matching pursuit, spatia l pyramid max pooling, and contrast normalization. We investigate the architectu re of HMP, and show that all three components are critical for good performance. To speed up the orthogonal matching pursuit, we propose a batch tree orthogonal matching pursuit that is particularly suitable to encode a large number of obse rvations that share the same large dictionary. HMP is scalable and can efficient ly handle full-size images. In addition, HMP enables linear support vector machi nes (SVM) to match the performance of nonlinear SVM while being scalable to larg e datasets. We compare HMP with many state-of-the-art algorithms including convo lutional deep belief networks, SIFT based single layer sparse coding, and kernel based feature learning. HMP consistently yields superior accuracy on three type s of image classification problems: object recognition (Caltech-101), scene reco gnition (MIT-Scene), and static event recognition (UIUC-Sports).
****
Solving Decision Problems with Limited Information
Denis D. Maua, Cassio Campos
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.
****
How Do Humans Teach: On Curriculum Learning and Teaching Dimension
Faisal Khan, Bilge Mutlu, Jerry Zhu
We study the empirical strategies that humans follow as they teach a target conc ept with a simple 1D threshold to a robot. Previous studies of computational te aching, particularly the teaching dimension model and the curriculum learning pr inciple, offer contradictory predictions on what optimal strategy the teacher sh ould follow in this teaching task. We show through behavioral studies that human

s employ three distinct teaching strategies, one of which is consistent with the curriculum learning principle, and propose a novel theoretical framework as a potential explanation for this strategy. This framework, which assumes a teaching goal of minimizing the learner's expected generalization error at each iteration, extends the standard teaching dimension model and offers a theoretical justification for curriculum learning.

****

A rational model of causal inference with continuous causes

Thomas Griffiths, Michael James

Rational models of causal induction have been successful in accounting for people's judgments about the existence of causal relationships. However, these models have focused on explaining inferences from discrete data of the kind that can be summarized in a 2 × 2 contingency table. This severely limits the scope of these models, since the world often provides non-binary data. We develop a new rational model of causal induction using continuous dimensions, which aims to diminish the gap between empirical and theoretical approaches and real-world causal induction. This model successfully predicts human judgments from previous studies better than models of discrete causal inference, and outperforms several other plausible models of causal induction with continuous causes in accounting for people's inferences in a new experiment.

****

Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis

Shuai Huang, Jing Li, Jieping Ye, Teresa Wu, Kewei Chen, Adam Fleisher, Eric Reiman

Diagnosis of Alzheimer's disease (AD) at the early stage of the disease development is of great clinical importance. Current clinical assessment that relies primarily on cognitive measures proves low sensitivity and specificity. The fast growing neuroimaging techniques hold great promise. Research so far has focused on single neuroimaging modalities. However, as different modalities provide complementary measures for the same disease pathology, fusion of multi-modality data may increase the statistical power in identification of disease-related brain regions. This is especially true for early AD, at which stage the disease-related regions are most likely to be weak-effect regions that are difficult to be detected from a single modality alone. We propose a sparse composite linear discriminant analysis model (SCLDA) for identification of disease-related brain regions of early AD from multi-modality data. SCLDA uses a novel formulation that decomposes each LDA parameter into a product of a common parameter shared by all the modalities and a parameter specific to each modality, which enables joint analysis of all the modalities and borrowing strength from one another. We prove that this formulation is equivalent to a penalized likelihood with non-convex regularization, which can be solved by the DC ((difference of convex functions) programming. We show that in using the DC programming, the property of the non-convex regularization in terms of preserving weak-effect features can be nicely revealed. We perform extensive simulations to show that SCLDA outperforms existing competing algorithms on feature selection, especially on the ability for identifying weak-effect features. We apply SCLDA to the Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) images of 49 AD patients and 67 normal controls (NC). Our study identifies disease-related brain regions consistent with findings in the AD literature.

****

Structured sparse coding via lateral inhibition

Arthur Szlam, Karol Gregor, Yann Cun

This work describes a conceptually simple method for structured sparse coding and dictionary design. Supposing a dictionary with K atoms, we introduce a structure as a set of penalties or interactions between every pair of atoms. We describe modifications of standard sparse coding algorithms for inference in this setting, and describe experiments showing that these algorithms are efficient. We show that interesting dictionaries can be learned for interactions that encode tree structures or locally connected structures. Finally, we show that our framework

allows us to learn the values of the interactions from the data, rather than having them pre-specified.
****

TD_gamma: Re-evaluating Complex Backups in Temporal Difference Learning
George Konidaris, Scott Niekum, Philip S. Thomas
We show that the lambda-return target used in the TD(lambda) family of algorithms is the maximum likelihood estimator for a specific model of how the variance of an n-step return estimate increases with n. We introduce the gamma-return estimator, an alternative target based on a more accurate model of variance, which defines the TDgamma family of complex-backup temporal difference learning algorithms. We derive TDgamma, the gamma-return equivalent of the original TD(lambda) algorithm, which eliminates the lambda parameter but can only perform updates at the end of an episode and requires time and space proportional to the episode length. We then derive a second algorithm, TDgamma(C), with a capacity parameter C. TDgamma(C) requires C times more time and memory than TD(lambda) and is incremental and online. We show that TDgamma outperforms TD(lambda) for any setting of lambda on 4 out of 5 benchmark domains, and that TDgamma(C) performs as well as or better than TD_gamma for intermediate settings of C.
****

Estimating time-varying input signals and ion channel states from a single voltage trace of a neuron
Ryota Kobayashi, Yasuhiro Tsubo, Petr Lansky, Shigeru Shinomoto
State-of-the-art statistical methods in neuroscience have enabled us to fit mathematical models to experimental data and subsequently to infer the dynamics of hidden parameters underlying the observable phenomena. Here, we develop a Bayesian method for inferring the time-varying mean and variance of the synaptic input, along with the dynamics of each ion channel from a single voltage trace of a neuron. An estimation problem may be formulated on the basis of the state-space model with prior distributions that penalize large fluctuations in these parameters. After optimizing the hyperparameters by maximizing the marginal likelihood, the state-space model provides the time-varying parameters of the input signals and the ion channel states. The proposed method is tested not only on the simulated data from the Hodgkin-Huxley type models but also on experimental data obtained from a cortical slice in vitro.
****

Joint 3D Estimation of Objects and Scene Layout
Andreas Geiger, Christian Wojek, Raquel Urtasun
We propose a novel generative model that is able to reason jointly about the 3D scene layout as well as the 3D location and orientation of objects in the scene. In particular, we infer the scene topology, geometry as well as traffic activities from a short video sequence acquired with a single camera mounted on a moving car. Our generative model takes advantage of dynamic information in the form of vehicle tracklets as well as static information coming from semantic labels and geometry (i.e., vanishing points). Experiments show that our approach outperforms a discriminative baseline based on multiple kernel learning (MKL) which has access to the same image information. Furthermore, as we reason about objects in 3D, we are able to significantly increase the performance of state-of-the-art object detectors in their ability to estimate object orientation.
****

Sparse Manifold Clustering and Embedding
Ehsan Elhamifar, René Vidal
We propose an algorithm called Sparse Manifold Clustering and Embedding (SMCE) for simultaneous clustering and dimensionality reduction of data lying in multiple nonlinear manifolds. Similar to most dimensionality reduction methods, SMCE finds a small neighborhood around each data point and connects each point to its neighbors with appropriate weights. The key difference is that SMCE finds both the neighbors and the weights automatically. This is done by solving a sparse optimization problem, which encourages selecting nearby points that lie in the same manifold and approximately span a low-dimensional affine subspace. The optimal solution encodes information that can be used for clustering and dimensionality r

eduction using spectral clustering and embedding. Moreover, the size of the optimal neighborhood of a data point, which can be different for different points, provides an estimate of the dimension of the manifold to which the point belongs. Experiments demonstrate that our method can effectively handle multiple manifolds that are very close to each other, manifolds with non-uniform sampling and holes, as well as estimate the intrinsic dimensions of the manifolds.
****

Submodular Multi-Label Learning
James Petterson, Tibério Caetano
In this paper we present an algorithm to learn a multi-label classifier which attempts at directly optimising the F-score. The key novelty of our formulation is that we explicitly allow for assortative (submodular) pairwise label interactions, i.e., we can leverage the co-ocurrence of pairs of labels in order to improve the quality of prediction. Prediction in this model consists of minimising a particular submodular set function, what can be accomplished exactly and efficiently via graph-cuts. Learning however is substantially more involved and requires the solution of an intractable combinatorial optimisation problem. We present an approximate algorithm for this problem and prove that it is sound in the sense that it never predicts incorrect labels. We also present a nontrivial test of a sufficient condition for our algorithm to have found an optimal solution. We present experiments on benchmark multi-label datasets, which attest the value of our proposed technique. We also make available source code that enables the reproduction of our experiments.
****

Learning Probabilistic Non-Linear Latent Variable Models for Tracking Complex Activities
Angela Yao, Juergen Gall, Luc V. Gool, Raquel Urtasun
A common approach for handling the complexity and inherent ambiguities of 3D human pose estimation is to use pose priors learned from training data. Existing approaches however, are either too simplistic (linear), too complex to learn, or can only learn latent spaces from "simple data", i.e., single activities such as walking or running. In this paper, we present an efficient stochastic gradient descent algorithm that is able to learn probabilistic non-linear latent spaces composed of multiple activities. Furthermore, we derive an incremental algorithm for the online setting which can update the latent space without extensive relearning. We demonstrate the effectiveness of our approach on the task of monocular and multi-view tracking and show that our approach outperforms the state-of-the-art.
****

Collective Graphical Models
Daniel R. Sheldon, Thomas Dietterich
There are many settings in which we wish to fit a model of the behavior of individuals but where our data consist only of aggregate information (counts or low-dimensional contingency tables). This paper introduces Collective Graphical Models---a framework for modeling and probabilistic inference that operates directly on the sufficient statistics of the individual model. We derive a highly-efficient Gibbs sampling algorithm for sampling from the posterior distribution of the sufficient statistics conditioned on noisy aggregate observations, prove its correctness, and demonstrate its effectiveness experimentally.
****

Sparse Estimation with Structured Dictionaries
David Wipf
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.
****

Newtron: an Efficient Bandit algorithm for Online Multiclass Prediction
Elad Hazan, Satyen Kale
We present an efficient algorithm for the problem of online multiclass predictio

n with bandit feedback in the fully adversarial setting. We measure its regret w ith respect to the log-loss defined in \cite{AbernethyR09}, which is parameteriz ed by a scalar (\alpha). We prove that the regret of \newtron is (O(\log T)) whe n (\alpha) is a constant that does not vary with horizon (T), and at most (O(T^{ 2/3})) if (\alpha) is allowed to increase to infinity with (T). For (\alpha) = (O(\log T)), the regret is bounded by (O(\sqrt{T})), thus solving the open proble m of \cite{KST08, AbernethyR09}. Our algorithm is based on a novel application o f the online Newton method \cite{HAK07}. We test our algorithm and show it to pe rform well in experiments, even when (\alpha) is a small constant.
****

The Fixed Points of Off-Policy TD
J. Kolter
Off-policy learning, the ability for an agent to learn about a policy other than the one it is following, is a key element of Reinforcement Learning, and in rec ent years there has been much work on developing Temporal Different (TD) algorit hms that are guaranteed to converge under off-policy sampling. It has remained an open question, however, whether anything can be said a priori about the quali ty of the TD solution when off-policy sampling is employed with function approxi mation. In general the answer is no: for arbitrary off-policy sampling the erro r of the TD solution can be unboundedly large, even when the approximator can re present the true value function well. In this paper we propose a novel approach to address this problem: we show that by considering a certain convex subset of off-policy distributions we can indeed provide guarantees as to the solution qu ality similar to the on-policy case. Furthermore, we show that we can efficient ly project on to this convex set using only samples generated from the system. The end result is a novel TD algorithm that has approximation guarantees even in the case of off-policy sampling and which empirically outperforms existing TD m ethods.
****

Transfer from Multiple MDPs
Alessandro Lazaric, Marcello Restelli
Transfer reinforcement learning (RL) methods leverage on the experience collecte d on a set of source tasks to speed-up RL algorithms. A simple and effective app roach is to transfer samples from source tasks and include them in the training set used to solve a target task. In this paper, we investigate the theoretical p roperties of this transfer method and we introduce novel algorithms adapting the transfer process on the basis of the similarity between source and target tasks . Finally, we report illustrative experimental results in a continuous chain pro blem.
****

Pylon Model for Semantic Segmentation
Victor Lempitsky, Andrea Vedaldi, Andrew Zisserman
Graph cut optimization is one of the standard workhorses of image segmentation s ince for binary random field representations of the image, it gives globally op timal results and there are efficient polynomial time implementations. Often, the random field is applied over a flat partitioning of the image into non-inte rsecting elements, such as pixels or super-pixels. In the paper we show that if, instead of a flat partitioning, the image is represented by a hierarchical segmentation tree, then the resulting energy combining unary and boundary terms can still be optimized using graph cut (with all the corresponding benefits of g lobal optimality and efficiency). As a result of such inference, the image gets partitioned into a set of segments that may come from different layers of the tr ee. We apply this formulation, which we call the pylon model, to the task of semantic segmentation where the goal is to separate an image into areas belongin g to different semantic classes. The experiments highlight the advantage of infe rence on a segmentation tree (over a flat partitioning) and demonstrate that the optimization in the pylon model is able to flexibly choose the level of segment ation across the image. Overall, the proposed system has superior segmentation a ccuracy on several datasets (Graz-02, Stanford background) compared to previousl y suggested approaches.

****
How biased are maximum entropy models?
Jakob H. Macke, Iain Murray, Peter Latham
Maximum entropy models have become popular statistical models in neuroscience and other areas in biology, and can be useful tools for obtaining estimates of mutual information in biological systems. However, maximum entropy models fit to small data sets can be subject to sampling bias; i.e. the true entropy of the data can be severely underestimated. Here we study the sampling properties of estimates of the entropy obtained from maximum entropy models. We show that if the data is generated by a distribution that lies in the model class, the bias is equal to the number of parameters divided by twice the number of observations. However, in practice, the true distribution is usually outside the model class, and we show here that this misspecification can lead to much larger bias. We provide a perturba- tive approximation of the maximally expected bias when the true model is out of model class, and we illustrate our results using numerical simulations of an Ising model; i.e. the second-order maximum entropy distribution on binary data.
****
Gaussian process modulated renewal processes
Yee Teh, Vinayak Rao
Renewal processes are generalizations of the Poisson process on the real line, whose intervals are drawn i.i.d. from some distribution. Modulated renewal processes allow these distributions to vary with time, allowing the introduction nonstationarity. In this work, we take a nonparametric Bayesian approach, modeling this nonstationarity with a Gaussian process. Our approach is based on the idea of uniformization, allowing us to draw exact samples from an otherwise intractable distribution. We develop a novel and efficient MCMC sampler for posterior inference. In our experiments, we test these on a number of synthetic and real datasets.
****
An ideal observer model for identifying the reference frame of objects
Joseph Austerweil, Abram L. Friesen, Thomas Griffiths
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.
****
Greedy Algorithms for Structurally Constrained High Dimensional Problems
Ambuj Tewari, Pradeep Ravikumar, Inderjit Dhillon
A hallmark of modern machine learning is its ability to deal with high dimensional problems by exploiting structural assumptions that limit the degrees of freedom in the underlying model. A deep understanding of the capabilities and limits of high dimensional learning methods under specific assumptions such as sparsity, group sparsity, and low rank has been attained. Efforts (Negahban et al., 2010, Chandrasekaran et al., 2010} are now underway to distill this valuable experience by proposing general unified frameworks that can achieve the twin goals of summarizing previous analyses and enabling their application to notions of structure hitherto unexplored. Inspired by these developments, we propose and analyze a general computational scheme based on a greedy strategy to solve convex optimization problems that arise when dealing with structurally constrained high-dimensional problems. Our framework not only unifies existing greedy algorithms by recovering them as special cases but also yields novel ones. Finally, we extend our results to infinite dimensional problems by using interesting connections between smoothness of norms and behavior of martingales in Banach spaces.
****