Sign in to GitHub · GitHub
****
An Optimal Policy for Target Localization with Application to Electron Microscop
y
Raphael Sznitman, Aurelien Lucchi, Peter Frazier, Bruno Jedynak, Pascal Fua
This paper considers the task of finding a target location by making a limited n
umber of sequential observations. Each observation results from evaluating an i
mperfect classifier of a chosen cost and accuracy on an interval of chosen lengt
h and position. Within a Bayesian framework, we study the problem of minimizing
 an objective that combines the entropy of the posterior distribution with the c
ost of the questions asked. In this problem, we show that the one-step lookahea
d policy is Bayes-optimal for any arbitrary time horizon. Moreover, this one-st
ep lookahead policy is easy to compute and implement. We then use this policy in
 the context of localizing mitochondria in electron microscope images, and exper
imentally show that significant speed ups in acquisition can be gained, while ma
intaining near equal image quality at target locations, when compared to current
 policies.
****
Domain Generalization via Invariant Feature Representation
Krikamol Muandet, David Balduzzi, Bernhard Schölkopf
This paper investigates domain generalization: How to take knowledge acquired fr
om an arbitrary number of related domains and apply it to previously unseen doma
ins? We propose Domain-Invariant Component Analysis (DICA), a kernel-based optim
ization algorithm that learns an invariant transformation by minimizing the diss
imilarity across domains, whilst preserving the functional relationship between
input and output variables. A learning-theoretic analysis shows that reducing di
ssimilarity improves the expected generalization ability of classifiers on new d
omains, motivating the proposed algorithm. Experimental results on synthetic and
 real-world datasets demonstrate that DICA successfully learns invariant feature
s and improves classifier performance in practice.
****
A Spectral Learning Approach to Range-Only SLAM
Byron Boots, Geoff Gordon
We present a novel spectral learning algorithm for simultaneous localization and
 mapping (SLAM) from range data with known correspondences. This algorithm is a
n instance of a general spectral system identification framework, from which it
inherits several desirable properties, including statistical consistency and no
local optima. Compared with popular batch optimization or multiple-hypothesis tr
acking (MHT) methods for range-only SLAM, our spectral approach offers guarantee
d low computational requirements and good tracking performance. Compared with MH
T and with popular extended Kalman filter (EKF) or extended information filter (
EIF) approaches, our approach does not need to linearize a transition or measure
ment model. We provide a theoretical analysis of our method, including finite-sa
mple error bounds. Finally, we demonstrate on a real-world robotic SLAM problem
 that our algorithm is not only theoretically justified, but works well in pract
ice: in a comparison of multiple methods, the lowest errors come from a combinat
ion of our algorithm with batch optimization, but our method alone produces near
ly as good a result at far lower computational cost.
****
Near-Optimal Bounds for Cross-Validation via Loss Stability
Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, Andrea Vattani
Multi-fold cross-validation is an established practice to estimate the error rat
e   of a learning algorithm. Quantifying the variance reduction gains due to cr
oss-validation   has been challenging due to the inherent correlations introduce
d by the folds.   In this work we introduce a new and weak measure of stability
 called \emphloss stability and relate the cross-validation performance to loss
 stability; we also establish that this   relationship is near-optimal. Our wor
k thus quantitatively improves the current best bounds on cross-validation.
****
Sparsity-Based Generalization Bounds for Predictive Sparse Coding

Nishant Mehta, Alexander Gray

The goal of predictive sparse coding is to learn a representation of examples as sparse linear combinations of elements from a dictionary, such that a learned hypothesis linear in the new representation performs well on a predictive task. Predictive sparse coding has demonstrated impressive performance on a variety of supervised tasks, but its generalization properties have not been studied. We establish the first generalization error bounds for predictive sparse coding, in the overcomplete setting, where the number of features k exceeds the original dimensionality d. The learning bound decays as (sqrt(d k/m)) with respect to d, k, and the size m of the training sample. It depends intimately on stability properties of the learned sparse encoder, as measured on the training sample. Consequently, we also present a fundamental stability result for the LASSO, a result that characterizes the stability of the sparse codes with respect to dictionary perturbations.

****

Sparse Uncorrelated Linear Discriminant Analysis

Xiaowei Zhang, Delin Chu

In this paper, we develop a novel approach for sparse uncorrelated linear discriminant analysis (ULDA). Our proposal is based on characterization of all solutions of the generalized ULDA. We incorporate sparsity into the ULDA transformation by seeking the solution with minimum \ell_1-norm from all minimum dimension solutions of the generalized ULDA. The problem is then formulated as a \ell_1-minimization problem and is solved by accelerated linearized Bregman method. Experiments on high-dimensional gene expression data demonstrate that our approach not only computes extremely sparse solutions but also performs well in classification. Experimental results also show that our approach can help for data visualization in low-dimensional space.

****

Block-Coordinate Frank-Wolfe Optimization for Structural SVMs

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, Patrick Pletscher

We propose a randomized block-coordinate variant of the classic Frank-Wolfe algorithm for convex optimization with block-separable constraints. Despite its lower iteration cost, we show that it achieves a similar convergence rate in duality gap as the full Frank-Wolfe algorithm. We also show that, when applied to the dual structural support vector machine (SVM) objective, this yields an online algorithm that has the same low iteration complexity as primal stochastic subgradient methods. However, unlike stochastic subgradient methods, the block-coordinate Frank-Wolfe algorithm allows us to compute the optimal step-size and yields a computable duality gap guarantee. Our experiments indicate that this simple algorithm outperforms competing structural SVM solvers.

****

Fast Probabilistic Optimization from Noisy Gradients

Philipp Hennig

Stochastic gradient descent remains popular in large-scale machine learning, on account of its very low computational cost and robustness to noise. However, gradient descent is only linearly efficient and not transformation invariant. Scaling by a local measure can substantially improve its performance. One natural choice of such a scale is the Hessian of the objective function: Were it available, it would turn linearly efficient gradient descent into the quadratically efficient Newton-Raphson optimization. Existing covariant methods, though, are either super-linearly expensive or do not address noise. Generalising recent results, this paper constructs a nonparametric Bayesian quasi-Newton algorithm that learns gradient and Hessian from noisy evaluations of the gradient. Importantly, the resulting algorithm, like stochastic gradient descent, has cost linear in the number of input dimensions.

****

Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes

Ohad Shamir, Tong Zhang

Stochastic Gradient Descent (SGD) is one of the simplest and most popular stocha

stic optimization methods. While it has already been theoretically studied for d
ecades, the classical analysis usually required non-trivial smoothness assumptio
ns, which do not apply to many modern applications of SGD with non-smooth object
ive functions such as support vector machines. In this paper, we investigate th
e performance of SGD \emphwithout such smoothness assumptions, as well as a runn
ing average scheme to convert the SGD iterates to a solution with optimal optimi
zation accuracy. In this framework, we prove that after T rounds, the suboptimal
ity of the \emphlast SGD iterate scales as O(\log(T)/\sqrtT) for non-smooth conv
ex objective functions, and O(\log(T)/T) in the non-smooth strongly convex case.
 To the best of our knowledge, these are the first bounds of this kind, and almo
st match the minimax-optimal rates obtainable by appropriate averaging schemes.
We also propose a new and simple averaging scheme, which not only attains optima
l rates, but can also be easily computed on-the-fly (in contrast, the suffix ave
raging scheme proposed in \citetRakhShaSri12arxiv is not as simple to implement)
. Finally, we provide some experimental illustrations.
****
Stochastic Alternating Direction Method of Multipliers
Hua Ouyang, Niao He, Long Tran, Alexander Gray
The Alternating Direction Method of Multipliers (ADMM) has received lots of atte
ntion recently due to the tremendous demand from large-scale and data-distribute
d machine learning applications. In this paper, we present a stochastic setting
for optimization problems with non-smooth composite objective functions. To solv
e this problem, we propose a stochastic ADMM algorithm. Our algorithm applies to
 a more general class of convex and nonsmooth objective functions, beyond the sm
ooth and separable least squares loss used in lasso. We also demonstrate the rat
es of convergence for our algorithm under various structural assumptions of the
stochastic function: O(1/\sqrtt) for convex functions and O(\log t/t) for strong
ly convex functions. Compared to previous literature, we establish the convergen
ce rate of ADMM for convex problems in terms of both the objective value and the
 feasibility violation. A novel application named Graph-Guided SVM is proposed t
o demonstrate the usefulness of our algorithm.
****
Noisy Sparse Subspace Clustering
Yu-Xiang Wang, Huan Xu
This paper considers the problem of subspace clustering under noise. Specificall
y, we study the behavior of Sparse Subspace Clustering (SSC) when either adversa
rial or random noise is added to the unlabelled input data points, which are ass
umed to lie in a union of low-dimensional subspaces. We show that a modified ve
rsion of SSC is \emphprovably effective in correctly identifying the underlying
subspaces, even with noisy data. This extends theoretical guarantee of this algo
rithm to the practical setting and provides justification to the success of SSC
in a class of real applications.
****
Parallel Markov Chain Monte Carlo for Nonparametric Mixture Models
Sinead Williamson, Avinava Dubey, Eric Xing
Nonparametric mixture models based on the Dirichlet process are an elegant alter
native to finite models when the number of underlying components is unknown, but
 inference in such models can be slow. Existing attempts to parallelize inferenc
e in such models have relied on introducing approximations, which can lead to in
accuracies in the posterior estimate. In this paper, we describe auxiliary varia
ble representations for the Dirichlet process and the hierarchical Dirichlet pro
cess that allow us to perform MCMC using the correct equilibrium distribution, i
n a distributed manner. We show that our approach allows scalable inference with
out the deterioration in estimate quality that accompanies existing methods.
****
Risk Bounds and Learning Algorithms for the Regression Approach to Structured Ou
tput Prediction
Sébastien Giguère, François Laviolette, Mario Marchand, Khadidja Sylla
We provide rigorous guarantees for the regression approach to structured output
prediction. We show that the quadratic regression loss is a convex surrogate of

the prediction loss when the output kernel satisfies some condition with respect to the prediction loss. We provide two upper bounds of the prediction risk that depend on the empirical quadratic risk of the predictor. The minimizer of the first bound is the predictor proposed by Cortes et al. (2007) while the minimizer of the second bound is a predictor that has never been proposed so far. Both predictors are compared on practical tasks.

****

Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures

James Bergstra, Daniel Yamins, David Cox

Many computer vision algorithms depend on configuration settings that are typically hand-tuned in the course of evaluating the algorithm for a particular data set. While such parameter tuning is often presented as being incidental to the algorithm, correctly setting these parameter choices is frequently critical to realizing a method's full potential. Compounding matters, these parameters often must be re-tuned when the algorithm is applied to a new problem domain, and the tuning process itself often depends on personal experience and intuition in ways that are hard to quantify or describe. Since the performance of a given technique depends on both the fundamental quality of the algorithm and the details of its tuning, it is sometimes difficult to know whether a given technique is genuinely better, or simply better tuned.   In this work, we propose a meta-modeling approach to support automated hyperparameter optimization, with the goal of providing practical tools that replace hand-tuning with a reproducible and unbiased optimization process. Our approach is to expose the underlying expression graph of how a performance metric (e.g. classification accuracy on validation examples) is computed from hyperparameters that govern not only how individual processing steps are applied, but even which processing steps are included. A hyperparameter optimization algorithm transforms this graph into a program for optimizing that performance metric. Our approach yields state of the art results on three disparate computer vision problems: a face-matching verification task (LFW), a face identification task (PubFig83) and an object recognition task (CIFAR-10), using a single broad class of feed-forward vision architectures.

****

Gibbs Max-Margin Topic Models with Fast Sampling Algorithms

Jun Zhu, Ning Chen, Hugh Perkins, Bo Zhang

Existing max-margin supervised topic models rely on an iterative procedure to solve multiple latent SVM subproblems with additional mean-field assumptions on the desired posterior distributions. This paper presents Gibbs max-margin supervised topic models by minimizing an expected margin loss, an upper bound of the existing margin loss derived from an expected prediction rule. By introducing augmented variables, we develop simple and fast Gibbs sampling algorithms with no restricting assumptions and no need to solve SVM subproblems for both classification and regression. Empirical results demonstrate significant improvements on time efficiency. The classification performance is also significantly improved over competitors.

****

Cost-Sensitive Tree of Classifiers

Zhixiang Xu, Matt Kusner, Kilian Weinberger, Minmin Chen

Recently, machine learning algorithms have successfully entered large-scale real-world industrial applications (e.g. search engines and email spam filters). Here, the CPU cost during test-time must be budgeted and accounted for. In this paper, we address the challenge of balancing test-time cost and the classifier accuracy in a principled fashion. The test-time cost of a classifier is often dominated by the computation required for feature extraction-which can vary drastically across features. We incorporate this extraction time by constructing a tree of classifiers, through which test inputs traverse along individual paths. Each path extracts different features and is optimized for a specific sub-partition of the input space. By only computing features for inputs that benefit from them the most, our cost-sensitive tree of classifiers can match the high accuracies of the current state-of-the-art at a small fraction of the computational cost.

****
Learning Hash Functions Using Column Generation
Xi Li, Guosheng Lin, Chunhua Shen, Anton Hengel, Anthony Dick
Fast nearest neighbor searching is becoming an increasingly important tool in solving many large-scale problems. Recently a number of approaches to learning data-dependent hash functions have been developed. In this work, we propose a column generation based method for learning data-dependent hash functions on the basis of proximity comparison information. Given a set of triplets that encode the pairwise proximity comparison information, our method learns hash functions that preserve the relative comparison relationships in the data as well as possible within the large-margin learning framework. The learning procedure is implemented using column generation and hence is named CGHash. At each iteration of the column generation procedure, the best hash function is selected. Unlike most other hashing methods, our method generalizes to new data points naturally; and has a training objective which is convex, thus ensuring that the global optimum can be identified. Experiments demonstrate that the proposed method learns compact binary codes and that its retrieval performance compares favorably with state-of-the-art methods when tested on a few benchmark datasets.
****
Combinatorial Multi-Armed Bandit: General Framework and Applications
Wei Chen, Yajun Wang, Yang Yuan
We define a general framework for a large class of combinatorial multi-armed bandit (CMAB) problems, where simple arms with unknown istributions form \em super arms. In each round, a super arm is played and the outcomes of its related simple arms are observed, which helps the selection of super arms in future rounds. The reward of the super arm depends on the outcomes of played arms, and it only needs to satisfy two mild assumptions, which allow a large class of nonlinear reward instances. We assume the availability of an (α,β)-approximation oracle that takes the means of the distributions of arms and outputs a super arm that with probability βgenerates an αfraction of the optimal expected reward. The objective of a CMAB algorithm is to minimize \em (α,β)-approximation regret, which is the difference in total expected reward between the $\alpha\beta$fraction of expected reward when always playing the optimal super arm, and the expected reward of playing super arms according to the algorithm. We provide CUCB algorithm that achieves O(\log n) regret, where n is the number of rounds played, and we further provide distribution-independent bounds for a large class of reward functions. Our regret analysis is tight in that it matches the bound for classical MAB problem up to a constant factor, and it significantly improves the regret bound in a recent paper on combinatorial bandits with linear rewards. We apply our CMAB framework to two new applications, probabilistic maximum coverage (PMC) for online advertising and social influence maximization for viral marketing, both having nonlinear reward structures.
****
Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization
Yuxin Chen, Andreas Krause
Active learning can lead to a dramatic reduction in labeling effort. However, in many practical implementations (such as crowdsourcing, surveys, high-throughput experimental design), it is preferable to query labels for batches of examples to be labelled in parallel. While several heuristics have been proposed for batch-mode active learning, little is known about their theoretical performance. We consider batch mode active learning and more general information-parallel stochastic optimization problems that exhibit adaptive submodularity, a natural diminishing returns condition. We prove that for such problems, a simple greedy strategy is competitive with the optimal batch-mode policy. In some cases, surprisingly, the use of batches incurs competitively low cost, even when compared to a fully sequential strategy. We demonstrate the effectiveness of our approach on batch-mode active learning tasks, where it outperforms the state of the art, as well as the novel problem of multi-stage influence maximization in social networks.
****

Convex formulations of radius-margin based Support Vector Machines
Huyen Do, Alexandros Kalousis
We consider Support Vector Machines (SVMs) learned together with linear transformations of the feature spaces on which they are applied. Under this scenario the radius of the smallest data enclosing sphere is no longer fixed. Therefore optimizing the SVM error bound by considering both the radius and the margin has the potential to deliver a tighter error bound. In this paper we present two novel algorithms: R-SVM$_\mu^+$—a SVM radius-margin based feature selection algorithm, and R-SVM$^+$ — a metric learning-based SVM. We derive our algorithms by exploiting a new tighter approximation of the radius and a metric learning interpretation of SVM. Both optimize directly the radius-margin error bound using linear transformations. Unlike almost all existing radius-margin based SVM algorithms which are either non-convex or combinatorial, our algorithms are standard quadratic convex optimization problems with linear or quadratic constraints. We perform a number of experiments on benchmark datasets. R-SVM$_\mu^+$ exhibits excellent feature selection performance compared to the state-of-the-art feature selection methods, such as L_1-norm and elastic-net based methods. R-SVM$^+$ achieves a significantly better classification performance compared to SVM and its other state-of-the-art variants. From the results it is clear that the incorporation of the radius, as a means to control the data spread, in the cost function has strong beneficial effects.
****
Modelling Sparse Dynamical Systems with Compressed Predictive State Representations
William L. Hamilton, Mahdi Milani Fard, Joelle Pineau
Efficiently learning accurate models of dynamical systems is of central importance for developing rational agents that can succeed in a wide range of challenging domains. The difficulty of this learning problem is particularly acute in settings with large observation spaces and partial observability. We present a new algorithm, called Compressed Predictive State Representation (CPSR), for learning models of high-dimensional partially observable uncontrolled dynamical systems from small sample sets. The algorithm, which extends previous work on Predictive State Representations, exploits a particular sparse structure present in many domains. This sparse structure is used to compress information during learning, allowing for an increase in both the efficiency and predictive power. The compression technique also relieves the burden of domain specific feature selection and allows for domains with extremely large discrete observation spaces to be efficiently modelled. We present empirical results showing that the algorithm is able to build accurate models more efficiently than its uncompressed counterparts, and provide theoretical results on the accuracy of the learned compressed model.
****
A Machine Learning Framework for Programming by Example
Aditya Menon, Omer Tamuz, Sumit Gulwani, Butler Lampson, Adam Kalai
Learning programs is a timely and interesting challenge. In Programming by Example (PBE), a system attempts to infer a program from input and output examples alone, by searching for a composition of some set of base functions. We show how machine learning can be used to speed up this seemingly hopeless search problem, by learning weights that relate textual features describing the provided input-output examples to plausible sub-components of a program. This generic learning framework lets us address problems beyond the scope of earlier PBE systems. Experiments on a prototype implementation show that learning improves search and ranking on a variety of text processing tasks found on help forums.
****
Discriminatively Activated Sparselets
Ross Girshick, Hyun Oh Song, Trevor Darrell
Shared representations are highly appealing due to their potential for gains in computational and statistical efficiency. Compressing a shared representation leads to greater computational savings, but at the same time can severely decrease performance on a target task. Recently, sparselets (Song et al., 2012) were introduced as a new shared intermediate representation for multiclass object

detection with deformable part models (Felzenszwalb et al., 2010a), showing si gnificant speedup factors, but with a large decrease in task performance. In t his paper we describe a new training framework that learns which sparselets to activate in order to optimize a discriminative objective, leading to larger spe edup factors with no decrease in task performance. We first reformulate sparse lets in a general structured output prediction framework, then analyze when sp arselets lead to computational efficiency gains, and lastly show experimental r esults on object detection and image classification tasks. Our experimental re sults demonstrate that discriminative activation substantially outperforms the previous reconstructive approach which, together with our structured output pre diction formulation, make sparselets broadly applicable and significantly more effective.
****
The Pairwise Piecewise-Linear Embedding for Efficient Non-Linear Classification
Ofir Pele, Ben Taskar, Amir Globerson, Michael Werman
Linear classiffers are much faster to learn and test than non-linear ones. On th e other hand, non-linear kernels offer improved performance, albeit at the incre ased cost of training kernel classiffers. To use non-linear mappings with effici ent linear learning algorithms, explicit embeddings that approximate popular ker nels have recently been proposed. However, the embedding process itself is often costly and the results are usually less accurate than kernel methods. In this w ork we propose a non-linear feature map that is both very efficient, but at the same time highly expressive. The method is based on discretization and interpola tion of individual features values and feature pairs. The discretization allows us to model different regions of the feature space separately, while the interpo lation preserves the original continuous values. Using this embedding is strictl y more general than a linear model and as efficient as the second-order polynomi al explicit feature map. An extensive empirical evaluation shows that our method consistently signiffcantly outperforms other methods, including a wide range of kernels. This is in contrast to other proposed embeddings that were faster than kernel methods, but with lower accuracy.
****
Fixed-Point Model For Structured Labeling
Quannan Li, Jingdong Wang, David Wipf, Zhuowen Tu
In this paper, we propose a simple but effective solution to the structured lab eling problem: a fixed-point model. Recently, layered models with sequential c lassifiers/regressors have gained an increasing amount of interests for struct ural prediction. Here, we design an algorithm with a new perspective on layered models; we aim to find a fixed-point function with the structured labels bein g both the output and the input. Our approach alleviates the burden in learnin g multiple/different classifiers in different layers. We devise a training str ategy for our method and provide justifications for the fixed-point function t o be a contraction mapping. The learned function captures rich contextual infor mation and is easy to train and test. On several widely used benchmark dataset s, the proposed method observes significant improvement in both performance an d efficiency over many state-of-the-art algorithms.
****
Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation
Boqing Gong, Kristen Grauman, Fei Sha
Learning domain-invariant features is of vital importance to unsupervised domain adaptation, where classifiers trained on the source domain need to be adapted t o a different target domain for which no labeled examples are available. In this paper, we propose a novel approach for learning such features. The central idea is to exploit the existence of landmarks, which are a subset of labeled data in stances in the source domain that are distributed most similarly to the target d omain. Our approach automatically discovers the landmarks and use them to bridge the source to the target by constructing provably easier auxiliary domain adapt ation tasks. The solutions of those auxiliary tasks form the basis to compose in variant features for the original task. We show how this composition can be opti

mized discriminatively without requiring labels from the target domain. We valid
ate the method on standard benchmark datasets for visual object recognition and
sentiment analysis of text. Empirical results show the proposed method outperfor
ms the state-of-the-art significantly.
****
Fast Conical Hull Algorithms for Near-separable Non-negative Matrix Factorizatio
n
Abhishek Kumar, Vikas Sindhwani, Prabhanjan Kambadur
The separability assumption (Arora et al., 2012; Donoho & Stodden, 2003) turns n
on-negative matrix factorization (NMF) into a tractable problem. Recently, a new
 class of provably-correct NMF algorithms have emerged under this assumption. In
 this paper, we reformulate the separable NMF problem as that of finding the ext
reme rays of the conical hull of a finite set of vectors. From this geometric pe
rspective, we derive new separable NMF algorithms that are highly scalable and e
mpirically noise robust, and have several favorable properties in relation to ex
isting methods. A parallel implementation of our algorithm scales excellently on
 shared and distributed-memory machines.
****
Principal Component Analysis on non-Gaussian Dependent Data
Fang Han, Han Liu
In this paper, we analyze the performance of a semiparametric principal componen
t analysis named Copula Component Analysis (COCA) (Han & Liu, 2012) when the dat
a are dependent. The semiparametric model assumes that, after unspecified margin
ally monotone transformations, the distributions are multivariate Gaussian. We s
tudy the scenario where the observations are drawn from non-i.i.d. processes ($m
$-dependency or a more general ϕ-mixing case). We show that COCA can allow
weak dependence. In particular, we provide the generalization bounds of converge
nce for both support recovery and parameter estimation of COCA for the dependent
 data. We provide explicit sufficient conditions on the degree of dependence, un
der which the parametric rate can be maintained. To our knowledge, this is the f
irst work analyzing the theoretical performance of PCA for the dependent data in
 high dimensional settings. Our results strictly generalize the analysis in Han
& Liu (2012) and the techniques we used have the separate interest for analyzing
 a variety of other multivariate statistical methods.
****
Learning Linear Bayesian Networks with Latent Variables
Animashree Anandkumar, Daniel Hsu, Adel Javanmard, Sham Kakade
This work considers the problem of learning linear Bayesian networks when some
of the variables are unobserved. Identifiability and efficient recovery from lo
w-order observable moments are established under a novel graphical constraint.
 The constraint concerns the expansion properties of the underlying directed ac
yclic graph (DAG) between observed and unobserved variables in the network, and
 it is satisfied by many natural families of DAGs that include multi-level DAGs
, DAGs with effective depth one, as well as certain families of polytrees.
****
Multiple Identifications in Multi-Armed Bandits
Séebastian Bubeck, Tengyao Wang, Nitin Viswanathan
We study the problem of identifying the top m arms in a multi-armed bandit game.
 Our proposed solution relies on a new algorithm based on successive rejects of
the seemingly bad arms, and successive accepts of the good ones. This algorithmi
c contribution allows to tackle other multiple identifications settings that wer
e previously out of reach. In particular we show that this idea of successive ac
cepts and rejects applies to the multi-bandit best arm identification problem.
****
Learning Optimally Sparse Support Vector Machines
Andrew Cotter, Shai Shalev-Shwartz, Nati Srebro
We show how to train SVMs with an optimal guarantee on the number of support vec
tors (up to constants), and with sample complexity and training runtime bounds m
atching the best known for kernel SVM optimization (i.e. without any additional
asymptotic cost beyond standard SVM training). Our method is simple to implement

and works well in practice.
****
Dynamic Probabilistic Models for Latent Feature Propagation in Social Networks
Creighton Heaukulani, Zoubin Ghahramani
Current Bayesian models for dynamic social network data have focused on modelling the influence of evolving unobserved structure on observed social interactions. However, an understanding of how observed social relationships from the past affect future unobserved structure in the network has been neglected. In this paper, we introduce a new probabilistic model for capturing this phenomenon, which we call latent feature propagation, in social networks. We demonstrate our model's capability for inferring such latent structure in varying types of social network datasets, and experimental studies show this structure achieves higher predictive performance on link prediction and forecasting tasks.
****
Efficient Sparse Group Feature Selection via Nonconvex Optimization
Shuo Xiang, Xiaoshen Tong, Jieping Ye
Sparse feature selection has been demonstrated to be effective in handling high-dimensional data. While promising, most of the existing works use convex methods, which may be suboptimal in terms of the accuracy of feature selection and parameter estimation. In this paper, we expand a nonconvex paradigm to sparse group feature selection, which is motivated by applications that require identifying the underlying group structure and performing feature selection simultaneously. The main contributions of this article are twofold: (1) computationally, we introduce a nonconvex sparse group feature selection model and present an efficient optimization algorithm, of which the key step is a projection with two coupled constraints; (2) statistically, we show that the proposed model can reconstruct the oracle estimator. Therefore, consistent feature selection and parameter estimation can be achieved. Numerical results on synthetic and real-world data suggest that the proposed nonconvex method compares favorably against its competitors, thus achieving desired goal of delivering high performance.
****
Domain Adaptation for Sequence Labeling Tasks with a Probabilistic Language Adaptation Model
Min Xiao, Yuhong Guo
In this paper, we propose to address the problem of domain adaptation for sequence labeling tasks via distributed representation learning by using a log-bilinear language adaptation model. The proposed neural probabilistic language model simultaneously models two different but related data distributions in the source and target domains   based on induced distributed representations, which encode both generalizable and domain-specific latent features. We then use the learned dense real-valued representation as   augmenting features for natural language processing systems. We empirically evaluate the proposed learning technique on WSJ and MEDLINE domains with POS tagging systems, and on WSJ and Brown corpora with syntactic chunking and name entity recognition systems. Our primary results show that the proposed domain adaptation method outperforms a number comparison methods for cross domain sequence labeling tasks.
****
Maximum Variance Correction with Application to A* Search
Wenlin Chen, Kilian Weinberger, Yixin Chen
In this paper we introduce Maximum Variance Correction (MVC), which finds large-scale feasible solutions to Maximum Variance Unfolding (MVU) by post-processing embeddings from any manifold learning algorithm. It increases the scale of MVU embeddings by several orders of magnitude and is naturally parallel. This unprecedented scalability opens up new avenues of applications for manifold learning, in particular the use of MVU embeddings as effective heuristics to speed-up A* search (Rayner et al. 2011).   We demonstrate that MVC embeddings lead to un-matched reductions in search time across several non-trivial A* benchmark search problems and bridge the gap between the manifold learning literature and one of its most promising high impact applications.
****

Adaptive Sparsity in Gaussian Graphical Models
Eleanor Wong, Suyash Awate, P. Thomas Fletcher
An effective approach to structure learning and parameter estimation for Gaussian graphical models is to impose a sparsity prior, such as a Laplace prior, on the entries of the precision matrix. Such an approach involves a hyperparameter that must be tuned to control the amount of sparsity. In this paper, we introduce a parameter-free method for estimating a precision matrix with sparsity that adapts to the data automatically. We achieve this by formulating a hierarchical Bayesian model of the precision matrix with a non-informative Jeffreys’ hyperprior. We also naturally enforce the symmetry and positive-definiteness constraints on the precision matrix by parameterizing it with the Cholesky decomposition. Experiments on simulated and real (cell signaling) data demonstrate that the proposed approach not only automatically adapts the sparsity of the model, but it also results in improved estimates of the precision matrix compared to the Laplace prior model with sparsity parameter chosen by cross-validation.
****
Average Reward Optimization Objective In Partially Observable Domains
Yuri Grinberg, Doina Precup
We consider the problem of average reward optimization in domains with partial observability, within the modeling framework of linear predictive state representations (PSRs). The key to average-reward computation is to have a well-defined stationary behavior of a system, so the required averages can be computed. If, additionally, the stationary behavior varies smoothly with changes in policy parameters, average-reward control through policy search also becomes a possibility. In this paper, we show that PSRs have a well-behaved stationary distribution, which is a rational function of policy parameters. Based on this result, we define a related reward process particularly suitable for average reward optimization, and analyze its properties. We show that in such a predictive state reward process, the average reward is a rational function of the policy parameters, whose complexity depends on the dimension of the underlying linear PSR. This result suggests that average reward-based policy search methods can be effective when the dimension of the system is small, even when the system representation in the POMDP framework requires many hidden states. We provide illustrative examples of this type.
****
Feature Selection in High-Dimensional Classification
Mladen Kolar, Han Liu
High-dimensional discriminant analysis is of fundamental importance in multivariate statistics. Existing theoretical results sharply characterize different procedures, providing sharp convergence results for the classification risk, as well as the l2 convergence results to the discriminative rule. However, sharp theoretical results for the problem of variable selection have not been established, even though model interpretation is of importance in many scientific domains. In this paper, we bridge this gap by providing sharp sufficient conditions for consistent variable selection using the ROAD estimator (Fan et al., 2010). Our results provide novel theoretical insights for the ROAD estimator. Sufficient conditions are complemented by the necessary information theoretic limits on variable selection in high-dimensional discriminant analysis. This complementary result also establishes optimality of the ROAD estimator for a certain family of problems.
****
Human Boosting
Harsh Pareek, Pradeep Ravikumar
Humans may be exceptional learners but they have biological limitations and moreover, inductive biases similar to machine learning algorithms. This puts limits on human learning ability and on the kinds of learning tasks humans can easily handle. In this paper, we consider the problem of “boosting” human learners to extend the learning ability of human learners and achieve improved performance on tasks which individual humans find difficult. We consider classification (category learning) tasks, propose a boosting algorithm for human learners and give the

oretical justifications. We conduct experiments using Amazon's Mechanical Turk on two synthetic datasets – a crosshair task with a nonlinear decision boundary and a gabor patch task with a linear boundary but which is inaccessible to human learners – and one real world dataset – the Opinion Spam detection task introduced in (Ott et al). Our results show that boosting human learners produces gains in accuracy and can overcome some fundamental limitations of human learners.
****

Efficient Dimensionality Reduction for Canonical Correlation Analysis

Haim Avron, Christos Boutsidis, Sivan Toledo, Anastasios Zouzias

We present a fast algorithm for approximate Canonical Correlation Analysis (CCA). Given a pair of tall-and-thin matrices, the proposed algorithm first employs a randomized dimensionality reduction transform to reduce the size of the input matrices, and then applies any standard CCA algorithm to the new pair of matrices. The algorithm computes an approximate CCA to the original pair of matrices with provable guarantees, while requiring asymptotically less operations than the state-of-the-art exact algorithms.
****

Parsing epileptic events using a Markov switching process model for correlated time series

Drausin Wulsin, Emily Fox, Brian Litt

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable numbers of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics. We demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.
****

Optimal rates for stochastic convex optimization under Tsybakov noise condition

Aaditya Ramdas, Aarti Singh

We focus on the problem of minimizing a convex function f over a convex set S given T queries to a stochastic first order oracle. We argue that the complexity of convex minimization is only determined by the rate of growth of the function around its minimum x^*_f,S, as quantified by a Tsybakov-like noise condition. Specifically, we prove that if f grows at least as fast as \|x-x^*_f,S\|^κaround its minimum, for some κ> 1, then the optimal rate of learning f(x^*_f,S) is Θ(T^-\fracκ2κ-2). The classic rate Θ(1/\sqrt T) for convex functions and Θ(1/T) for strongly convex functions are special cases of our result for κ→∞and κ=2, and even faster rates are attained for 1 < κ< 2. We also derive tight bounds for the complexity of learning x_f,S^*, where the optimal rate is Θ(T^-\frac12κ-2). Interestingly, these precise rates also characterize the complexity of active learning and our results further strengthen the connections between the fields of active learning and convex optimization, both of which rely on feedback-driven queries.
****

A Randomized Mirror Descent Algorithm for Large Scale Multiple Kernel Learning

Arash Afkanpour, András György, Csaba Szepesvari, Michael Bowling

We consider the problem of simultaneously learning to linearly combine a very large number of kernels and learn a good predictor based on the learnt kernel. When the number of kernels d to be combined is very large, multiple kernel learning methods whose computational cost scales linearly in d are intractable. We propo

se a randomized version of the mirror descent algorithm to overcome this issue, under the objective of minimizing the group p-norm penalized empirical risk. The key to achieve the required exponential speed-up is the computationally efficient construction of low-variance estimates of the gradient. We propose importance sampling based estimates, and find that the ideal distribution samples a coordinate with a probability proportional to the magnitude of the corresponding gradient. We show that in the case of learning the coefficients of a polynomial kernel, the combinatorial structure of the base kernels to be combined allows sampling from this distribution in O(\log(d)) time, making the total computational cost of the method to achieve an epsilon-optimal solution to be O(\log(d)/epsilon^2), thereby allowing our method to operate for very large values of d. Experiments with simulated and real data confirm that the new algorithm is computationally more efficient than its state-of-the-art alternatives.
****
Noisy and Missing Data Regression: Distribution-Oblivious Support Recovery
Yudong Chen, Constantine Caramanis
Many models for sparse regression typically assume that the covariates are known completely, and without noise. Particularly in high-dimensional applications, this is often not the case. Worse yet, even estimating statistics of the noise (the noise covariance) can be a central challenge. In this paper we develop a simple variant of orthogonal matching pursuit (OMP) for precisely this setting. We show that without knowledge of the noise covariance, our algorithm recovers the support, and we provide matching lower bounds that show that our algorithm performs at the minimax optimal rate. While simple, this is the first algorithm that (provably) recovers support in a noise-distribution-oblivious manner. When knowledge of the noise-covariance is available, our algorithm matches the best-known \ell^2-recovery bounds available. We show that these too are min-max optimal. Along the way, we also obtain improved performance guarantees for OMP for the standard sparse regression problem with Gaussian noise.
****
Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method
Taiji Suzuki
We develop new stochastic optimization methods that are applicable to   a wide range of structured regularizations. Basically our methods are combinations of basic stochastic optimization techniques and Alternating Direction Multiplier Method (ADMM). ADMM is a general framework for optimizing a composite function, and has a wide range of applications. We propose two types of online variants of ADMM,   which correspond to online proximal gradient descent and regularized dual averaging respectively. The proposed algorithms are computationally efficient and easy to implement. Our methods yield O(1/\sqrtT) convergence of the expected risk. Moreover, the online proximal gradient descent type method yields O(\log(T)/T) convergence for a strongly convex loss. Numerical experiments show effectiveness of our methods in learning tasks with structured sparsity such as overlapped group lasso.
****
A New Frontier of Kernel Design for Structured Data
Kilho Shin
Many kernels for discretely structured data in the literature are designed within the framework of the convolution kernel and its generalization, the mapping kernel. The two most important advantages to use this framework is an easy-to-check criteria of positive definiteness and efficient computation based on the dynamic programming methodology of the resulting kernels. On the other hand, the recent theory of partitionable kernels reveals that the known kernels only take advantage of a very small portion of the potential of the framework. In fact, we have good opportunities to find novel and important kernels in the unexplored area. In this paper, we shed light on a novel important class of kernels within the framework: We give a mathematical characterization of the class, show a parametric method to optimize kernels of the class to specific problems, based on this characterization, and present some experimental results, which show the new ker

nels are promising in both accuracy and efficiency.
****
Learning with Marginalized Corrupted Features
Laurens Maaten, Minmin Chen, Stephen Tyree, Kilian Weinberger
The goal of machine learning is to develop predictors that generalize well to test data. Ideally, this is achieved by training on very large (infinite) training data sets that capture all variations in the data distribution. In the case of finite training data, an effective solution is to extend the training set with artificially created examples – which, however, is also computationally costly. We propose to corrupt training examples with noise from known distributions within the exponential family and present a novel learning algorithm, called marginalized corrupted features (MCF), that trains robust predictors by minimizing the expected value of the loss function under the corrupting distribution – essentially learning with infinitely many (corrupted) training examples. We show empirically on a variety of data sets that MCF classifiers can be trained efficiently, may generalize substantially better to test data, and are more robust to feature deletion at test time.
****
Approximation properties of DBNs with binary hidden units and real-valued visible units
Oswin Krause, Asja Fischer, Tobias Glasmachers, Christian Igel
Deep belief networks (DBNs) can approximate any distribution over fixed-length binary vectors. However, DBNs are frequently applied to model real-valued data, and so far little is known about their representational power in this case. We analyze the approximation properties of DBNs with two layers of binary hidden units and visible units with conditional distributions from the exponential family. It is shown that these DBNs can, under mild assumptions, model any additive mixture of distributions from the exponential family with independent variables. An arbitrarily good approximation in terms of Kullback-Leibler divergence of an m-dimensional mixture distribution with n components can be achieved by a DBN with m visible variables and n and n+1 hidden variables in the first and second hidden layer, respectively. Furthermore, relevant infinite mixtures can be approximated arbitrarily well by a DBN with a finite number of neurons. This includes the important special case of an infinite mixture of Gaussian distributions with fixed variance restricted to a compact domain, which in turn can approximate any strictly positive density over this domain.
****
Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization
Martin Jaggi
We provide stronger and more general primal-dual convergence results for Frank-Wolfe-type algorithms (a.k.a. conditional gradient) for constrained convex optimization, enabled by a simple framework of duality gap certificates. Our analysis also holds if the linear subproblems are only solved approximately (as well as if the gradients are inexact), and is proven to be worst-case optimal in the sparsity of the obtained solutions.   On the application side, this allows us to unify a large variety of existing sparse greedy methods, in particular for optimization over convex hulls of an atomic set, even if those sets can only be approximated, including sparse (or structured sparse) vectors or matrices, low-rank matrices, permutation matrices, or max-norm bounded matrices.   We present a new general framework for convex optimization over matrix factorizations, where every Frank-Wolfe iteration will consist of a low-rank update, and discuss the broad application areas of this approach.
****
General Functional Matrix Factorization Using Gradient Boosting
Tianqi Chen, Hang Li, Qiang Yang, Yong Yu
Matrix factorization is among the most successful techniques for collaborative filtering. One challenge of collaborative filtering is how to utilize available auxiliary information to improve prediction accuracy. In this paper, we study the problem of utilizing auxiliary information as features of   factorization and propose formalizing the problem as general functional matrix factorization, whos

e model includes conventional matrix factorization models as its special cases. Moreover, we propose a gradient boosting based algorithm to efficiently solve the optimization problem. Finally, we give two specific algorithms for efficient feature function construction for two specific tasks. Our method can construct more suitable feature functions by searching in an infinite functional space based on training data and thus can yield better prediction accuracy. The experimental results demonstrate that the proposed method outperforms the baseline methods on three real-world datasets.
****
Iterative Learning and Denoising in Convolutional Neural Associative Memories
Amin Karbasi, Amir Hesam Salavati, Amin Shokrollahi
The task of a neural associative memory is to retrieve a set of previously memorized patterns from their noisy versions by using a network of neurons. Hence, an ideal network should be able to 1) gradually learn a set of patterns, 2) retrieve the correct pattern from noisy queries and 3) maximize the number of memorized patterns while maintaining the reliability in responding to queries. We show that by considering the inherent redundancy in the memorized patterns, one can obtain all the mentioned properties at once. This is in sharp contrast with the previous work that could only improve one or two aspects at the expense of the third. More specifically, we devise an iterative algorithm that learns the redundancy among the patterns. The resulting network has a retrieval capacity that is exponential in the size of the network. Lastly, by considering the local structures of the network, the asymptotic error correction performance can be made linear in the size of the network.
****
Scaling Multidimensional Gaussian Processes using Projected Additive Approximations
Elad Gilboa, Yunus Saatçi, John Cunningham, Elad Gilboa
Exact Gaussian Process (GP) regression has O(N^3) runtime for data size N, making it intractable for large N. Advances in GP scaling have not been extended to the multidimensional input setting, despite the preponderance of multidimensional applications.   This paper introduces and tests a novel method of projected additive approximation to multidimensional GPs. We thoroughly illustrate the power of this method on several datasets, achieving close performance to the naive Full GP at orders of magnitude less cost.
****
Active Learning for Multi-Objective Optimization
Marcela Zuluaga, Guillaume Sergent, Andreas Krause, Markus Püschel
In many fields one encounters the challenge of identifying, out of a pool of possible designs, those that simultaneously optimize multiple objectives. This means that usually there is not one optimal design but an entire set of Pareto-optimal ones with optimal tradeoffs in the objectives. In many applications, evaluating one design is expensive; thus, an exhaustive search for the Pareto-optimal set is unfeasible. To address this challenge, we propose the Pareto Active Learning (PAL) algorithm, which intelligently samples the design space to predict the Pareto-optimal set. Key features of PAL include (1) modeling the objectives as samples from a Gaussian process distribution to capture structure and accommodate noisy evaluation; (2) a method to carefully choose the next design to evaluate to maximize progress; and (3) the ability to control prediction accuracy and sampling cost. We provide theoretical bounds on PAL’s sampling cost required to achieve a desired accuracy. Further, we show an experimental evaluation on three real-world data sets. The results show PAL’s effectiveness; in particular it improves significantly over a state-of-the-art evolutionary algorithm, saving in many cases about 33%.
****
A Generalized Kernel Approach to Structured Output Learning
Hachem Kadri, Mohammad Ghavamzadeh, Philippe Preux
We study the problem of structured output learning from a regression perspective. We first provide a general formulation of the kernel dependency estimation (KDE) approach to this problem using operator-valued kernels. Our formulation overc

omes the two main limitations of the original KDE approach, namely the decoupling between outputs in the image space and the inability to use a joint feature space. We then propose a covariance-based operator-valued kernel that allows us to take into account the structure of the kernel feature space. This kernel operates on the output space and only encodes the interactions between the outputs without any reference to the input space. To address this issue, we introduce a variant of our KDE method based on the conditional covariance operator that in addition to the correlation between the outputs takes into account the effects of the input variables. Finally, we evaluate the performance of our KDE approach using both covariance and conditional covariance kernels on three structured output problems, and compare it to the state-of-the art kernel-based structured output regression methods.
****
Efficient Active Learning of Halfspaces: an Aggressive Approach
Alon Gonen, Sivan Sabato, Shai Shalev-Shwartz
We study pool-based active learning of half-spaces. We revisit the aggressive approach for active learning in the realizable case, and show that it can be made efficient and practical, while also having theoretical guarantees under reasonable assumptions. We further show, both theoretically and experimentally, that it can be preferable to mellow approaches. Our efficient aggressive active learner of half-spaces has formal approximation guarantees that hold when the pool is separable with a margin. While our analysis is focused on the realizable setting, we show that a simple heuristic allows using the same algorithm successfully for pools with low error as well. We further compare the aggressive approach to the mellow approach, and prove that there are cases in which the aggressive approach results in significantly better label complexity compared to the mellow approach. We demonstrate experimentally that substantial improvements in label complexity can be achieved using the aggressive approach, for both realizable and low-error settings.
****
Enhanced statistical rankings via targeted data collection
Braxton Osting, Christoph Brune, Stanley Osher
Given a graph where vertices represent alternatives and pairwise comparison data, y_ij, is given on the edges, the statistical ranking problem is to find a potential function, defined on the vertices, such that the gradient of the potential function agrees with pairwise comparisons. We study the dependence of the statistical ranking problem on the available pairwise data, i.e., pairs (i,j) for which the pairwise comparison data y_ij is known, and propose a framework to identify data which, when augmented with the current dataset, maximally increases the Fisher information of the ranking. Under certain assumptions, the data collection problem decouples, reducing to a problem of finding an edge set on the graph (with a fixed number of edges) such that the second eigenvalue of the graph Laplacian is maximal. This reduction of the data collection problem to a spectral graph-theoretic question is one of the primary contributions of this work. As an application, we study the Yahoo! Movie user rating dataset and demonstrate that the addition of a small number of well-chosen pairwise comparisons can significantly increase the Fisher informativeness of the ranking.
****
Online Feature Selection for Model-based Reinforcement Learning
Trung Nguyen, Zhuoru Li, Tomi Silander, Tze Yun Leong
We propose a new framework for learning the world dynamics of feature-rich environments in model-based reinforcement learning. The main idea is formalized as a new, factored state-transition representation that supports efficient online-learning of the relevant features. We construct the transition models through predicting how the actions change the world. We introduce an online sparse coding learning technique for feature selection in high-dimensional spaces. We derive theoretical guarantees for our framework and empirically demonstrate its practicality in both simulated and real robotics domains.
****
ELLA: An Efficient Lifelong Learning Algorithm

Paul Ruvolo, Eric Eaton
The problem of learning multiple consecutive tasks, known as lifelong learning, is of great importance to the creation of intelligent, general-purpose, and flexible machines. In this paper, we develop a method for online multi-task learning in the lifelong learning setting. The proposed Efficient Lifelong Learning Algorithm (ELLA) maintains a sparsely shared basis for all task models, transfers knowledge from the basis to learn each new task, and refines the basis over time to maximize performance across all tasks. We show that ELLA has strong connections to both online dictionary learning for sparse coding and state-of-the-art batch multi-task learning methods, and provide robust theoretical performance guarantees. We show empirically that ELLA yields nearly identical performance to batch multi-task learning while learning tasks sequentially in three orders of magnitude (over 1,000x) less time.
****
A Structural SVM Based Approach for Optimizing Partial AUC
Harikrishna Narasimhan, Shivani Agarwal
The area under the ROC curve (AUC) is a widely used performance measure in machine learning. Increasingly, however, in several applications, ranging from ranking and biometric screening to medical diagnosis, performance is measured not in terms of the full area under the ROC curve, but instead, in terms of the partial area under the ROC curve between two specified false positive rates. In this paper, we develop a structural SVM framework for directly optimizing the partial AUC between any two false positive rates. Our approach makes use of a cutting plane solver along the lines of the structural SVM based approach for optimizing the full AUC developed by Joachims (2005). Unlike the full AUC, where the combinatorial optimization problem needed to find the most violated constraint in the cutting plane solver can be decomposed easily to yield an efficient algorithm, the corresponding optimization problem in the case of partial AUC is harder to decompose. One of our key technical contributions is an efficient algorithm for solving this combinatorial optimization problem that has the same computational complexity as Joachims’ algorithm for optimizing the usual AUC. This allows us to efficiently optimize the partial AUC in any desired false positive range. We demonstrate the approach on a variety of real-world tasks.
****
Convex Relaxations for Learning Bounded-Treewidth Decomposable Graphs
K. S. Sesh Kumar, Francis Bach
We consider the problem of learning the structure of undirected graphical models with bounded treewidth, within the maximum likelihood framework. This is an NP-hard problem and most approaches consider local search techniques. In this paper, we pose it as a combinatorial optimization problem, which is then relaxed to a convex optimization problem that involves searching over the forest and hyperforest polytopes with special structures. A supergradient method is used to solve the dual problem, with a run-time complexity of O(k^3 n^k+2 \log n) for each iteration, where n is the number of variables and k is a bound on the treewidth. We compare our approach to state-of-the-art methods on synthetic datasets and classical benchmarks, showing the gains of the novel convex approach.
****
Adaptive Task Assignment for Crowdsourced Classification
Chien-Ju Ho, Shahin Jabbari, Jennifer Wortman Vaughan
Crowdsourcing markets have gained popularity as a tool for inexpensively collecting data from diverse populations of workers. Classification tasks, in which workers provide labels (such as “offensive” or “not offensive”) for instances (such as websites), are among the most common tasks posted, but due to a mix of human error and the overwhelming prevalence of spam, the labels collected are often noisy. This problem is typically addressed by collecting labels for each instance from multiple workers and combining them in a clever way. However, the question of how to choose which tasks to assign to each worker is often overlooked. We investigate the problem of task assignment and label inference for heterogeneous classification tasks. By applying online primal-dual techniques, we derive a provably near-optimal adaptive assignment algorithm. We show that adaptively assign

ing workers to tasks can lead to more accurate predictions at a lower cost when
the available workers are diverse.
****
Optimal Regret Bounds for Selecting the State Representation in Reinforcement L
earning
Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, Daniil Ryabko
We consider an agent interacting with an environment in a single stream of actio
ns, observations, and rewards, with no reset. This process is not assumed to be
a Markov Decision Process (MDP). Rather, the agent has several representations (
mapping histories of past interactions to a discrete state space) of the environ
ment with unknown dynamics, only some of which result in an MDP. The goal is to
minimize the average regret criterion against an agent who knows   an MDP repres
entation giving the highest optimal reward, and acts optimally in it. Recent reg
ret bounds for this setting are of order O(T^2/3) with an additive term constant
 yet exponential in some characteristics of the optimal MDP. We propose an algo
rithm whose regret after T time steps is O(\sqrtT), with all constants reasonab
ly small. This is optimal in T since O(\sqrtT) is the optimal regret in the set
ting of learning in a (single discrete) MDP.
****
Better Mixing via Deep Representations
Yoshua Bengio, Gregoire Mesnil, Yann Dauphin, Salah Rifai
It has been hypothesized, and supported with experimental evidence, that deeper
representations, when well trained, tend to do a better job at disentangling the
 underlying factors of variation. We study the following related conjecture: be
tter representations, in the sense of better disentangling, can be exploited to
produce Markov chains that mix faster between modes. Consequently, mixing betwee
n modes would be more efficient at higher levels of representation. To better u
nderstand this, we propose a secondary conjecture: the higher-level samples fill
 more uniformly the space they occupy and the high-density manifolds tend to unf
old when represented at higher levels. The paper discusses these hypotheses and
 tests them experimentally through visualization and measurements of mixing betw
een modes and interpolating between samples.
****
Online Latent Dirichlet Allocation with Infinite Vocabulary
Ke Zhai, Jordan Boyd-Graber
Topic models based on latent Dirichlet allocation (LDA) assume a predefined voca
bulary a priori. This is reasonable in batch settings, but it is not reasonable
when data are revealed over time, as is the case with streaming / online algorit
hms. To address this lacuna, we extend LDA by drawing topics from a Dirichlet pr
ocess whose base distribution is a distribution over all strings rather than fro
m a finite Dirichlet. We develop inference using online variational inference an
d because we only can consider a finite number of words for each truncated topic
 propose heuristics to dynamically organize, expand, and contract the set of wor
ds we consider in our vocabulary truncation. We show our model can successfully
incorporate new words as it encounters new terms and that it performs better tha
n online LDA in evaluations of topic quality and classification performance.
****
Characterizing the Representer Theorem
Yaoliang Yu, Hao Cheng, Dale Schuurmans, Csaba Szepesvari
The representer theorem assures that kernel methods retain optimality under pena
lized empirical risk minimization. While a sufficient condition on the form of t
he regularizer guaranteeing the representer theorem has been known since the ini
tial development of kernel methods, necessary conditions have only been investig
ated recently. In this paper we completely characterize the necessary and suffic
ient conditions on the regularizer that ensure the representer theorem holds. Th
e results are surprisingly simple yet broaden the conditions where the represent
er theorem is known to hold. Extension to the matrix domain is also addressed.
****
Dynamical Models and tracking regret in online convex programming
Eric Hall, Rebecca Willett

This paper describes a new online convex optimization method which incorporates a family of candidate dynamical models and establishes novel tracking regret bounds that scale with comparator's deviation from the best dynamical model in this family. Previous online optimization methods are designed to have a total accumulated loss comparable to that of the best comparator sequence, and existing tracking or shifting regret bounds scale with the overall variation of the comparator sequence. In many practical scenarios, however, the environment is nonstationary and comparator sequences with small variation are quite weak, resulting in large losses. The proposed dynamic mirror descent method, in contrast, can yield low regret relative to highly variable comparator sequences by both tracking the best dynamical model and forming predictions based on that model. This concept is demonstrated empirically in the context of sequential compressive observations of a dynamic scene and tracking a dynamic social network.
****
Large-Scale Bandit Problems and KWIK Learning
Jacob Abernethy, Kareem Amin, Michael Kearns, Moez Draief
We show that parametric multi-armed bandit (MAB) problems with large state and action spaces can be algorithmically reduced to the supervised learning model known as Knows What It Knows or KWIK learning. We give matching impossibility results showing that the KWIK learnability requirement cannot be replaced by weaker supervised learning assumptions. We provide such results in both the standard parametric MAB setting, as well as for a new model in which the action space is finite but growing with time.
****
Vanishing Component Analysis
Roi Livni, David Lehavi, Sagi Schein, Hila Nachliely, Shai Shalev-Shwartz, Amir Globerson
The vanishing ideal of a set of n points S, is the set of all polynomials that attain the value of zero on all the points in S. Such ideals can be compactly represented using a small set of polynomials known as generators of the ideal. Here we describe and analyze an efficient procedure that constructs a set of generators of a vanishing ideal. Our procedure is numerically stable, and can be used to find approximately vanishing polynomials. The resulting polynomials capture nonlinear structure in data, and can for example be used within supervised learning. Empirical comparison with kernel methods show that our method constructs more compact classifiers with comparable accuracy.
****
Learning an Internal Dynamics Model from Control Demonstration
Matthew Golub, Steven Chase, Byron Yu
Much work in optimal control and inverse control has assumed that the controller has perfect knowledge of plant dynamics. However, if the controller is a human or animal subject, the subject's internal dynamics model may differ from the true plant dynamics. Here, we consider the problem of learning the subject's internal model from demonstrations of control and knowledge of task goals. Due to sensory feedback delay, the subject uses an internal model to generate an internal prediction of the current plant state, which may differ from the actual plant state. We develop a probabilistic framework and exact EM algorithm to jointly estimate the internal model, internal state trajectories, and feedback delay. We applied this framework to demonstrations by a nonhuman primate of brain-machine interface (BMI) control. We discovered that the subject's internal model deviated from the true BMI plant dynamics and provided significantly better explanation of the recorded neural control signals than did the true plant dynamics.
****
Robust Structural Metric Learning
Daryl Lim, Gert Lanckriet, Brian McFee
Metric learning algorithms produce a linear transformation of data which is optimized for a prediction task, such as nearest-neighbor classification or ranking. However, when the input data contains a large portion of non-informative features, existing methods fail to identify the relevant features, and performance degrades accordingly. In this paper, we present an efficient and robust structural

metric learning algorithm which enforces group sparsity on the learned transformation, while optimizing for structured ranking output prediction. Experiments on synthetic and real datasets demonstrate that the proposed method outperforms previous methods in both high- and low-noise settings.
****
Constrained fractional set programs and their application in local clustering and community detection
Thomas Bühler, Shyam Sundar Rangapuram, Simon Setzer, Matthias Hein
The (constrained) minimization of a ratio of set functions is a problem frequently occurring in clustering and community detection. As these optimization problems are typically NP-hard, one uses convex or spectral relaxations in practice. While these relaxations can be solved globally optimally, they are often too loose and thus lead to results far away from the optimum. In this paper we show that every constrained minimization problem of a ratio of non-negative set functions allows a tight relaxation into an unconstrained continuous optimization problem. This result leads to a flexible framework for solving constrained problems in network analysis. While a globally optimal solution for the resulting non-convex problem cannot be guaranteed, we outperform the loose convex or spectral relaxations by a large margin on constrained local clustering problems.
****
Efficient Semi-supervised and Active Learning of Disjunctions
Nina Balcan, Christopher Berlind, Steven Ehrlich, Yingyu Liang
We provide efficient algorithms for learning disjunctions in the semi-supervised setting under a natural regularity assumption introduced by (Balcan & Blum, 2005). We prove bounds on the sample complexity of our algorithms under a mild restriction on the data distribution. We also give an active learning algorithm with improved sample complexity and extend all our algorithms to the random classification noise setting.
****
Convex Adversarial Collective Classification
MohamadAli Torkamani, Daniel Lowd
In this paper, we present a novel method for robustly performing collective classification in the presence of a malicious adversary that can modify up to a fixed number of binary-valued attributes. Our method is formulated as a convex quadratic program that guarantees optimal weights against a worst-case adversary in polynomial time. In addition to increased robustness against active adversaries, this kind of adversarial regularization can also lead to improved generalization even when no adversary is present. In experiments on real and simulated data, our method consistently outperforms both non-adversarial and non-relational baselines.
****
Rounding Methods for Discrete Linear Classification
Yann Chevaleyre, Frédéerick Koriche, Jean-daniel Zucker
Learning discrete linear functions is a notoriously difficult challenge. In this paper, the learning task is cast as combinatorial optimization problem: given a set of positive and negative feature vectors in the Euclidean space, the goal is to find a discrete linear function that minimizes the cumulative hinge loss of this training set. Since this problem is NP-hard, we propose two simple rounding algorithms that discretize the fractional solution of the problem. Generalization bounds are derived for two important classes of binary-weighted linear functions, by establishing the Rademacher complexity of these classes and proving approximation bounds for rounding methods. These methods are compared on both synthetic and real-world data.
****
Mixture of Mutually Exciting Processes for Viral Diffusion
Shuang-Hong Yang, Hongyuan Zha
\emphDiffusion network inference and \emphmeme tracking have been two key challenges in viral diffusion. This paper shows that these two tasks can be addressed simultaneously with a probabilistic model involving a mixture of mutually exciting point processes. A fast learning algorithms is developed based on mean-field

variational inference with budgeted diffusion bandwidth. The model is demonstrated with applications to the diffusion of viral texts in (1) online social networks (e.g., Twitter) and (2) the blogosphere on the Web.
****
Gaussian Process Vine Copulas for Multivariate Dependence
David Lopez-Paz, Jose Miguel Hernández-Lobato, Ghahramani Zoubin
Copulas allow to learn marginal distributions separately from the multivariate dependence structure (copula) that links them together into a density function. Vine factorizations ease the learning of high-dimensional copulas by constructing a hierarchy of conditional bivariate copulas. However, to simplify inference, it is common to assume that each of these conditional bivariate copulas is independent from its conditioning variables. In this paper, we relax this assumption by discovering the latent functions that specify the shape of a conditional copula given its conditioning variables. We learn these functions by following a Bayesian approach based on sparse Gaussian processes with expectation propagation for scalable, approximate inference. Experiments on real-world datasets show that, when modeling all conditional dependencies, we obtain better estimates of the underlying copula of the data.
****
Stochastic Simultaneous Optimistic Optimization
Michal Valko, Alexandra Carpentier, Rémi Munos
We study the problem of global maximization of a function f given a finite number of evaluations perturbed by noise. We consider a very weak assumption on the function, namely that it is locally smooth (in some precise sense) with respect to some semi-metric, around one of its global maxima. Compared to previous works on bandits in general spaces (Kleinberg et al., 2008; Bubeck et al., 2011a) our algorithm does not require the knowledge of this semi-metric. Our algorithm, StoSOO, follows an optimistic strategy to iteratively construct upper confidence bounds over the hierarchical partitions of the function domain to decide which point to sample next. A finite-time analysis of StoSOO shows that it performs almost as well as the best specifically-tuned algorithms even though the local smoothness of the function is not known.
****
Toward Optimal Stratification for Stratified Monte-Carlo Integration
Alexandra Carpentier, Rémi Munos
We consider the problem of adaptive stratified sampling for Monte Carlo integration of a function, given a finite number of function evaluations perturbed by noise. Here we address the problem of adapting simultaneously the number of samples into each stratum and the stratification itself. We show a tradeoff in the size of the partitioning. On the one hand it is important to refine the partition in areas where the observation noise or the function are heterogeneous in order to reduce this variability. But on the other hand, a too refined stratification makes it harder to assign the samples according to a near-optimal (oracle) allocation strategy. In this paper we provide an algorithm \em Monte-Carlo Upper-Lower Confidence Bound that selects online, among a large class of partitions, the partition that provides a near-optimal trade-off, and allocates the samples almost optimally on this partition.
****
A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems
Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, Jieping Ye
Non-convex sparsity-inducing penalties have recently received considerable attentions in sparse learning. Recent theoretical investigations have demonstrated their superiority over the convex counterparts in several sparse learning settings. However, solving the non-convex optimization problems associated with non-convex penalties remains a big challenge. A commonly used approach is the Multi-Stage (MS) convex relaxation (or DC programming), which relaxes the original non-convex problem to a sequence of convex problems. This approach is usually not very practical for large-scale problems because its computational cost is a multiple of solving a single convex problem. In this paper, we propose a General Iterativ

e Shrinkage and Thresholding (GIST) algorithm to solve the nonconvex optimization problem for a large class of non-convex penalties. The GIST algorithm iteratively solves a proximal operator problem, which in turn has a closed-form solution for many commonly used penalties. At each outer iteration of the algorithm, we use a line search initialized by the Barzilai-Borwein (BB) rule that allows finding an appropriate step size quickly. The paper also presents a detailed convergence analysis of the GIST algorithm. The efficiency of the proposed algorithm is demonstrated by extensive experiments on large-scale data sets.
****

Thurstonian Boltzmann Machines: Learning from Multiple Inequalities
Truyen Tran, Dinh Phung, Svetha Venkatesh
We introduce Thurstonian Boltzmann Machines (TBM), a unified architecture that can naturally incorporate a wide range of data inputs at the same time. Our motivation rests in the Thurstonian view that many discrete data types can be considered as being generated from a subset of underlying latent continuous variables, and in the observation that each realisation of a discrete type imposes certain inequalities on those variables. Thus learning and inference in TBM reduce to making sense of a set of inequalities. Our proposed TBM naturally supports the following types: Gaussian, intervals, censored, binary, categorical, muticategorical, ordinal, (in)-complete rank with and without ties. We demonstrate the versatility and capacity of the proposed model on three applications of very different natures; namely handwritten digit recognition, collaborative filtering and complex social survey analysis.
****

A Variational Approximation for Topic Modeling of Hierarchical Corpora
Do-kyum Kim, Geoffrey Voelker, Lawrence Saul
We study the problem of topic modeling in corpora whose documents are organized in a multi-level hierarchy. We explore a parametric approach to this problem, assuming that the number of topics is known or can be estimated by cross-validation. The models we consider can be viewed as special (finite-dimensional) instances of hierarchical Dirichlet processes (HDPs). For these models we show that there exists a simple variational approximation for probabilistic inference. The approximation relies on a previously unexploited inequality that handles the conditional dependence between Dirichlet latent variables in adjacent levels of the model’s hierarchy. We compare our approach to existing implementations of nonparametric HDPs. On several benchmarks we find that our approach is faster than Gibbs sampling and able to learn more predictive models than existing variational methods. Finally, we demonstrate the large-scale viability of our approach on two newly available corpora from researchers in computer security–one with 350,000 documents and over 6,000 internal subcategories, the other with a five-level deep hierarchy.
****

Forecastable Component Analysis
Georg Goerg
I introduce Forecastable Component Analysis (ForeCA), a novel dimension reduction technique for temporally dependent signals. Based on a new forecastability measure, ForeCA finds an optimal transformation to separate a multivariate time series into a forecastable and an orthogonal white noise space. I present a converging algorithm with a fast eigenvector solution. Applications to financial and macro-economic time series show that ForeCA can successfully discover informative structure, which can be used for forecasting as well as classification. The R package ForeCA accompanies this work and is publicly available on CRAN.
****

Ellipsoidal Multiple Instance Learning
Gabriel Krummenacher, Cheng Soon Ong, Joachim Buhmann
We propose a large margin method for asymmetric learning with ellipsoids, called eMIL, suited to multiple instance learning (MIL). We derive the distance between ellipsoids and the hyperplane, generalising the standard support vector machine. Negative bags in MIL contain only negative instances, and we treat them akin to uncertain observations in the robust optimisation framework. However, our met

hod allows positive bags to cross the margin, since it is not known which instan
ces within are positive. We show that representing bags as ellipsoids under the
 introduced distance is the most robust solution when treating a bag as a random
 variable with finite mean and covariance. Two algorithms are derived to solve t
he resulting non-convex optimization problem: a concave-convex procedure and a q
uasi-Newton method. Our method achieves competitive results on benchmark dataset
s. We introduce a MIL dataset from a real world application of detecting wheel d
efects from multiple partial observations, and show that eMIL outperforms compet
ing approaches.
****
Local Low-Rank Matrix Approximation
Joonseok Lee, Seungyeon Kim, Guy Lebanon, Yoram Singer
Matrix approximation is a common tool in recommendation systems, text mining, an
d computer vision. A prevalent assumption in constructing matrix approximations
is that the partially observed matrix is of low-rank. We propose a new matrix ap
proximation model where we assume instead that the matrix is locally of low-rank
, leading to a representation of the observed matrix as a weighted sum of low-ra
nk matrices. We analyze the accuracy of the proposed local low-rank modeling. Ou
r experiments show improvements in prediction accuracy over classical approaches
 for recommendation tasks.
****
Generic Exploration and K-armed Voting Bandits
Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, Sami Naamane
We study a stochastic online learning scheme with partial feedback where the uti
lity of decisions is only observable through an estimation of the environment pa
rameters. We propose a generic pure-exploration algorithm, able to cope with var
ious utility functions from multi-armed bandits settings to dueling bandits. The
 primary application of this setting is to offer a natural generalization of due
ling bandits for situations where the environment parameters reflect the idiosyn
cratic preferences of a mixed crowd.
****
A unifying framework for vector-valued manifold regularization and multi-view le
arning
Minh Hà Quang, Loris Bazzani, Vittorio Murino
This paper presents a general vector-valued reproducing kernel Hilbert spaces (R
KHS) formulation for the problem of learning an unknown functional dependency b
etween a structured input space and a structured output space, in the Semi-Super
vised Learning setting. Our formulation includes as special cases Vector-valued
Manifold Regularization and Multi-view Learning, thus provides in particular a u
nifying framework linking these two important learning approaches. In the case
of least square loss function, we provide a closed form solution with an efficie
nt implementation. Numerical experiments on challenging multi-class categorizati
on problems show that our multi-view learning formulation achieves results which
 are comparable with state of the art and are significantly better than single-v
iew learning.
****
Learning Connections in Financial Time Series
Gartheeban Ganeshapillai, John Guttag, Andrew Lo
To reduce risk, investors seek assets that have high expected return and are unl
ikely to move in tandem. Correlation measures are generally used to quantify the
 connections between equities. The 2008 financial crisis, and its aftermath, dem
onstrated the need for a better way to quantify these connections. We present a
machine learning-based method to build a connectedness matrix to address the sho
rtcomings of correlation in capturing events such as large losses. Our method us
es an unconstrained optimization to learn this matrix, while ensuring that the r
esulting matrix is positive semi-definite. We show that this matrix can be used
to build portfolios that not only “beat the market,” but also outperform optima
l (i.e., minimum variance) portfolios.
****
Fast dropout training

Sida Wang, Christopher Manning
Preventing feature co-adaptation by encouraging independent contributions from different features often improves classification and regression performance. Dropout training (Hinton et al., 2012) does this by randomly dropping out (zeroing) hidden units and input features during training of neural networks. However, repeatedly sampling a random subset of input features makes training much slower. Based on an examination of the implied objective function of dropout training, we show how to do fast dropout training by sampling from or integrating a Gaussian approximation, instead of doing Monte Carlo optimization of this objective. This approximation, justified by the central limit theorem and empirical evidence, gives an order of magnitude speedup and more stability. We show how to do fast dropout training for classification, regression, and multilayer neural networks. Beyond dropout, our technique is extended to integrate out other types of noise and small image transformations.
****
Scalable Optimization of Neighbor Embedding for Visualization
Zhirong Yang, Jaakko Peltonen, Samuel Kaski
Neighbor embedding (NE) methods have found their use in data visualization but are limited in big data analysis tasks due to their O(n^2) complexity for n data samples. We demonstrate that the obvious approach of subsampling produces inferior results and propose a generic approximated optimization technique that reduces the NE optimization cost to O(n log n). The technique is based on realizing that in visualization the embedding space is necessarily very low-dimensional (2D or 3D), and hence efficient approximations developed for n-body force calculations can be applied. In gradient-based NE algorithms the gradient for an individual point decomposes into “forces” exerted by the other points. The contributions of close-by points need to be computed individually but far-away points can be approximated by their “center of mass”, rapidly computable by applying a recursive decomposition of the visualization space into quadrants. The new algorithm brings a significant speed-up for medium-size data, and brings “big data” within reach of visualization.
****
Precision-recall space to correct external indices for biclustering
Blaise Hanczar, Mohamed Nadif
Biclustering is a major tool of data mining in many domains and many algorithms have emerged in recent years. All these algorithms aim to obtain coherent biclusters and it is crucial to have a reliable procedure for their validation. We point out the problem of size bias in biclustering evaluation and show how it can lead to wrong conclusions in a comparative study. We present the theoretical corrections for all of the most popular measures in order to remove this bias. We introduce the corrected precision-recall space that combines the advantages of corrected measures, the ease of interpretation and visualization of uncorrected measures. Numerical experiments demonstrate the interest of our approach.
****
Monochromatic Bi-Clustering
Sharon Wulff, Ruth Urner, Shai Ben-David
We propose a natural cost function for the bi-clustering task, the monochromatic cost. This cost function is suitable for detecting meaningful homogeneous bi-clusters based on categorical valued input matrices. Such tasks arise in many applications, such as the analysis of social networks and in systems-biology where researchers try to infer functional grouping of biological agents based on their pairwise interactions. We analyze the computational complexity of the resulting optimization problem. We present a polynomial time approximation algorithm for this bi-clustering task and complement this result by showing that finding (exact) optimal solutions is NP-hard. As far as we know, these are the first positive approximation guarantees and formal NP-hardness results for any bi-clustering optimization problem. In addition, we show that our optimization problem can be efficiently solved by deterministic annealing, yielding a promising heuristic for large problem instances.
****

Gated Autoencoders with Tied Input Weights
Droniou Alain, Sigaud Olivier
The semantic interpretation of images is one of the core applications of deep learning. Several techniques have been recently proposed to model the relation between two images, with application to pose estimation, action recognition or invariant object recognition. Among these techniques, higher-order Boltzmann machines or relational autoencoders consider projections of the images on different subspaces and intermediate layers act as transformation specific detectors. In this work, we extend the mathematical study of (Memisevic, 2012b) to show that it is possible to use a unique projection for both images in a way that turns intermediate layers as spectrum encoders of transformations. We show that this results in networks that are easier to tune and have greater generalization capabilities.
****
Strict Monotonicity of Sum of Squares Error and Normalized Cut in the Lattice of Clusterings
Nicola Rebagliati
Sum of Squares Error and Normalized Cut are two widely used clustering functional. It is known their minimum values are monotone with respect to the input number of clusters and this monotonicity does not allow for a simple automatic selection of a correct number of clusters. Here we study monotonicity not just on the minimizers but on the entire clustering lattice. We show the value of Sum of Squares Error is strictly monotone under the strict refinement relation of clusterings and we obtain data-dependent bounds on the difference between the value of a clustering and one of its refinements. Using analogous techniques we show the value of Normalized Cut is strictly anti-monotone. These results imply that even if we restrict our solutions to form a chain of clustering, like the one we get from hierarchical algorithms, we cannot rely on the functional values in order to choose the number of clusters. By using these results we get some data-dependent bounds on the difference of the values of any two clusterings.
****
Transition Matrix Estimation in High Dimensional Time Series
Fang Han, Han Liu
In this paper, we propose a new method in estimating transition matrices of high dimensional vector autoregressive (VAR) models. Here the data are assumed to come from a stationary Gaussian VAR time series. By formulating the problem as a linear program, we provide a new approach to conduct inference on such models. In theory, under a doubly asymptotic framework in which both the sample size T and dimensionality d of the time series can increase, we provide explicit rates of convergence between the estimator and the population transition matrix under different matrix norms. Our results show that the spectral norm of the transition matrix plays a pivotal role in determining the final rates of convergence. This is the first work analyzing the estimation of transition matrices under a high dimensional doubly asymptotic framework. Experiments are conducted on both synthetic and real-world stock data to demonstrate the effectiveness of the proposed method compared with the existing methods. The results of this paper have broad impact on different applications, including finance, genomics, and brain imaging.
****
Label Partitioning For Sublinear Ranking
Jason Weston, Ameesh Makadia, Hector Yee
We consider the case of ranking a very large set of labels, items, or documents, which is common to information retrieval, recommendation, and large-scale annotation tasks. We present a general approach for converting an algorithm which has linear time in the size of the set to a sublinear one via label partitioning. Our method consists of learning an input partition and a label assignment to each partition of the space such that precision at k is optimized, which is the loss function of interest in this setting. Experiments on large-scale ranking and recommendation tasks show that our method not only makes the original linear time algorithm computationally tractable, but can also improve its performance.
****

Subproblem-Tree Calibration: A Unified Approach to Max-Product Message Passing
Huayan Wang, Koller Daphne
Max-product (max-sum) message passing algorithms are widely used for MAP inference in MRFs. It has many variants sharing a common flavor of passing "messages" over some graph-object. Recent advances revealed that its convergent versions (such as MPLP, MSD, TRW-S) can be viewed as performing block coordinate descent (BCD) in a dual objective. That is, each BCD step achieves dual-optimal w.r.t. a block of dual variables (messages), thereby decreases the dual objective monotonically. However, most existing algorithms are limited to updating blocks selected in rather restricted ways. In this paper, we show a "unified" message passing algorithm that: (a) subsumes MPLP, MSD, and TRW-S as special cases when applied to their respective choices of dual objective and blocks, and (b) is able to perform BCD under much more flexible choices of blocks (including very large blocks) as well as the dual objective itself (that arise from an arbitrary dual decomposition).
****
Collaborative hyperparameter tuning
Rémi Bardenet, Mátyás Brendel, Balázs Kégl, Michèle Sebag
Hyperparameter learning has traditionally been a manual task because of the limited number of trials. Today’s computing infrastructures allow bigger evaluation budgets, thus opening the way for algorithmic approaches. Recently, surrogate-based optimization was successfully applied to hyperparameter learning for deep belief networks and to WEKA classifiers. The methods combined brute force computational power with model building about the behavior of the error function in the hyperparameter space, and they could significantly improve on manual hyperparameter tuning. What may make experienced practitioners even better at hyperparameter optimization is their ability to generalize across similar learning problems. In this paper, we propose a generic method to incorporate knowledge from previous experiments when simultaneously tuning a learning algorithm on new problems at hand. To this end, we combine surrogate-based ranking and optimization techniques for surrogate-based collaborative tuning (SCoT). We demonstrate SCoT in two experiments where it outperforms standard tuning techniques and single-problem surrogate-based optimization.
****
SADA: A General Framework to Support Robust Causation Discovery
Ruichu Cai, Zhenjie Zhang, Zhifeng Hao
Causality discovery without manipulation is considered a crucial problem to a variety of applications, such as genetic therapy. The state-of-the-art solutions, e.g. LiNGAM, return accurate results when the number of labeled samples is larger than the number of variables. These approaches are thus applicable only when large numbers of samples are available or the problem domain is sufficiently small. Motivated by the observations of the local sparsity properties on causal structures, we propose a general Split-and-Merge strategy, named SADA, to enhance the scalability of a wide class of causality discovery algorithms. SADA is able to accurately identify the causal variables, even when the sample size is significantly smaller than the number of variables. In SADA, the variables are partitioned into subsets, by finding cuts on the sparse probabilistic graphical model over the variables. By running mainstream causation discovery algorithms, e.g. LiNGAM, on the subproblems, complete causality can be reconstructed by combining all the partial results. SADA benefits from the recursive division technique, since each small subproblem generates more accurate result under the same number of samples. We theoretically prove that SADA always reduces the scale of problems without significant sacrifice on result accuracy, depending only on the local sparsity condition over the variables. Experiments on real-world datasets verify the improvements on scalability and accuracy by applying SADA on top of existing causation algorithms.
****
Learning and Selecting Features Jointly with Point-wise Gated Boltzmann Machines
Kihyuk Sohn, Guanyu Zhou, Chansoo Lee, Honglak Lee
Unsupervised feature learning has emerged as a promising tool in learning repres

entations from unlabeled data. However, it is still challenging to learn useful
high-level features when the data contains a significant amount of irrelevant pa
tterns. Although feature selection can be used for such complex data, it may fai
l when we have to build a learning system from scratch (i.e., starting from the
lack of useful raw features). To address this problem, we propose a point-wise g
ated Boltzmann machine, a unified generative model that combines feature learnin
g and feature selection. Our model performs not only feature selection on learne
d high-level features (i.e., hidden units), but also dynamic feature selection o
n raw features (i.e., visible units) through a gating mechanism. For each exampl
e, the model can adaptively focus on a variable subset of visible nodes correspo
nding to the task-relevant patterns, while ignoring the visible units correspond
ing to the task-irrelevant patterns. In experiments, our method achieves improve
d performance over state-of-the-art in several visual recognition benchmarks.
****

Sequential Bayesian Search
Zheng Wen, Branislav Kveton, Brian Eriksson, Sandilya Bhamidipati
Millions of people search daily for movies, music, and books on the Internet. Un
fortunately, non-personalized exploration of items can result in an infeasible n
umber of costly interaction steps. We study the problem of efficient, repeated i
nteractive search. In this problem, the user is navigated to the items of intere
st through a series of options and our objective is to learn a better search pol
icy from past interactions with the user. We propose an efficient learning algor
ithm for solving the problem, sequential Bayesian search (SBS), and prove that i
t is Bayesian optimal. We also analyze the algorithm from the frequentist point
of view and show that its regret is sublinear in the number of searches. Finally
, we evaluate our method on a real-world movie discovery problem and show that i
t performs nearly optimally as the number of searches increases.
****

Sparse projections onto the simplex
Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, Christoph Koch
Most learning methods with rank or sparsity constraints use convex relaxations,
which lead to optimization with the nuclear norm or the \ell_1-norm. However, se
veral important learning applications cannot benefit from this approach as they
feature these convex norms as constraints in addition to the non-convex rank and
 sparsity constraints. In this setting, we derive efficient sparse projections o
nto the simplex and its extension, and illustrate how to use them to solve high-
dimensional learning problems in quantum tomography, sparse density estimation a
nd portfolio selection with non-convex constraints.
****

Modeling Musical Influence with Topic Models
Uri Shalit, Daphna Weinshall, Gal Chechik
The role of musical influence has long been debated by scholars and critics in
the humanities, but never in a data-driven way. In this work we approach the qu
estion of influence by applying topic-modeling tools (Blei & Lafferty, 2006; Ger
rish & Blei, 2010) to a dataset of 24941 songs by 9222 artists, from the years 1
922 to 2010. We find the models to be significantly correlated with a human-cura
ted influence measure, and to clearly outperform a baseline method. Further usin
g the learned model to study properties of influence, we find that musical influ
ence and musical innovation are not monotonically correlated. However, we do fin
d that the most influential songs were more innovative during two time periods:
the early 1970’s and the mid 1990’s.
****

Subtle Topic Models and Discovering Subtly Manifested Software Concerns Automati
cally
Mrinal Das, Suparna Bhattacharya, Chiranjib Bhattacharyya, Gopinath Kanchi
In a recent pioneering approach LDA was used to discover cross cutting concerns(
CCC) automatically from software codebases. LDA though successful in detecting p
rominent concerns, fails to detect many useful CCCs including ones that may be h
eavily executed but elude discovery because they do not have a strong prevalence
 in source-code. We pose this problem as that of discovering topics that rarely

occur in individual documents, which we will refer to as subtle topics. Recently an interesting approach, namely focused topic models(FTM) was proposed for detecting rare topics. FTM, though successful in detecting topics which occur prominently in very few documents, is unable to detect subtle topics. Discovering subtle topics thus remains an important open problem. To address this issue we propose subtle topic models(STM). STM uses a generalized stick breaking process(GSBP) as a prior for defining multiple distributions over topics. This hierarchical structure on topics allows STM to discover rare topics beyond the capabilities of FTM. The associated inference is non-standard and is solved by exploiting the relationship between GSBP and generalized Dirichlet distribution. Empirical results show that STM is able to discover subtle CCC in two benchmark code-bases, a feat which is beyond the scope of existing topic models, thus demonstrating the potential of the model in automated concern discovery, a known difficult problem in Software Engineering. Furthermore it is observed that even in general text corpora STM outperforms the state of art in discovering subtle topics.
****

Exploring the Mind: Integrating Questionnaires and fMRI

Esther Salazar, Ryan Bogdan, Adam Gorka, Ahmad Hariri, Lawrence Carin

A new model is developed for joint analysis of ordered, categorical, real and count data. The ordered and categorical data are answers to questionnaires, the (word) count data correspond to the text questions from the questionnaires, and the real data correspond to fMRI responses for each subject. The Bayesian model employs the von Mises distribution in a novel manner to infer sparse graphical models jointly across people, questions, fMRI stimuli and brain region, with this integrated within a new matrix factorization based on latent binary features. The model is compared with simpler alternatives on two real datasets. We also demonstrate the ability to predict the response of the brain to visual stimuli (as measured by fMRI), based on knowledge of how the associated person answered classical questionnaires.
****

A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions

Quoc Tran Dinh, Anastasios Kyrillidis, Volkan Cevher

We propose an algorithmic framework for convex minimization problems of composite functions with two terms: a self-concordant part and a possibly nonsmooth regularization part. Our method is a new proximal Newton algorithm with local quadratic convergence rate. As a specific problem instance, we consider sparse precision matrix estimation problems in graph learning. Via a careful dual formulation and a novel analytic step-size selection, we instantiate an algorithm within our framework for graph learning that avoids Cholesky decompositions and matrix inversions, making it attractive for parallel and distributed implementations.
****

A Practical Algorithm for Topic Modeling with Provable Guarantees

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, Michael Zhu

Topic models provide a useful method for dimensionality reduction and exploratory data analysis in large text corpora. Most approaches to topic model learning have been based on a maximum likelihood objective. Efficient algorithms exist that attempt to approximate this objective, but they have no provable guarantees. Recently, algorithms have been introduced that provide provable bounds, but these algorithms are not practical because they are inefficient and not robust to violations of model assumptions. In this paper we present an algorithm for learning topic models that is both provable and practical. The algorithm produces results comparable to the best MCMC implementations while running orders of magnitude faster.
****

Distributed training of Large-scale Logistic models

Siddharth Gopal, Yiming Yang

Regularized Multinomial Logistic regression has emerged as one of the most common methods for performing data classification and analysis. With the advent of la

rge-scale data it is common to find scenarios where the number of possible multinomial outcomes is large (in the order of thousands to tens of thousands). In such cases, the computational cost of training logistic models or even simply iterating through all the model parameters is prohibitively expensive. In this paper, we propose a training method for large-scale multinomial logistic models that breaks this bottleneck by enabling parallel optimization of the likelihood objective. Our experiments on large-scale datasets showed an order of magnitude reduction in training time.
****
An Adaptive Learning Rate for Stochastic Variational Inference
Rajesh Ranganath, Chong Wang, Blei David, Eric Xing
Stochastic variational inference finds good posterior approximations of probabilistic models with very large data sets. It optimizes the variational objective with stochastic optimization, following noisy estimates of the natural gradient. Operationally, stochastic inference iteratively subsamples from the data, analyzes the subsample, and updates parameters with a decreasing learning rate. However, the algorithm is sensitive to that rate, which usually requires hand-tuning to each application. We solve this problem by developing an adaptive learning rate for stochastic inference. Our method requires no tuning and is easily implemented with computations already made in the algorithm. We demonstrate our approach with latent Dirichlet allocation applied to three large text corpora. Inference with the adaptive learning rate converges faster and to a better approximation than the best settings of hand-tuned rates.
****
Margins, Shrinkage, and Boosting
Matus Telgarsky
This manuscript shows that AdaBoost and its immediate variants can produce approximately maximum margin classifiers simply by scaling their step size choices by a fixed small constant. In this way, when the unscaled step size is an optimal choice, these results provide guarantees for Friedman's empirically successful "shrinkage" procedure for gradient boosting (Friedman, 2000). Guarantees are also provided for a variety of other step sizes, affirming the intuition that increasingly regularized line searches provide improved margin guarantees. The results hold for the exponential loss and similar losses, most notably the logistic loss.
****
Canonical Correlation Analysis based on Hilbert-Schmidt Independence Criterion and Centered Kernel Target Alignment
Billy Chang, Uwe Kruger, Rafal Kustra, Junping Zhang
Canonical correlation analysis (CCA) is a well established technique for identifying linear relationships among two variable sets. Kernel CCA (KCCA) is the most notable nonlinear extension but it lacks interpretability and robustness against irrelevant features. The aim of this article is to introduce two nonlinear CCA extensions that rely on the recently proposed Hilbert-Schmidt independence criterion and the centered kernel target alignment. These extensions determine linear projections that provide maximally dependent projected data pairs. The paper demonstrates that the use of linear projections allows removing irrelevant features, whilst extracting combinations of strongly associated features. This is exemplified through a simulation and the analysis of recorded data that are available in the literature.
****
Large-Scale Learning with Less RAM via Randomization
Daniel Golovin, D. Sculley, Brendan McMahan, Michael Young
We reduce the memory footprint of popular large-scale online learning methods by projecting our weight vector onto a coarse discrete set using randomized rounding. Compared to standard 32-bit float encodings, this reduces RAM usage by more than 50% during training and by up 95% when making predictions from a fixed model, with almost no loss in accuracy. We also show that randomized counting can be used to implement per-coordinate learning rates, improving model quality with little additional RAM. We prove these memory-saving methods achieve regret guaran

tees similar to their exact variants. Empirical evaluation confirms excellent pe rformance, dominating standard approaches across memory versus accuracy tradeoff s.
****
Taming the Curse of Dimensionality: Discrete Integration by Hashing and Optimiza tion
Stefano Ermon, Carla Gomes, Ashish Sabharwal, Bart Selman
Integration is affected by the curse of dimensionality and quickly becomes intra ctable as the dimensionality of the problem grows. We propose a randomized algor ithm that, with high probability, gives a constant-factor approximation of a gen eral discrete integral defined over an exponentially large set. This algorithm r elies on solving only a small number of instances of a discrete combinatorial op timization problem subject to randomly generated parity constraints used as a ha sh function. As an application, we demonstrate that with a small number of MAP q ueries we can efficiently approximate the partition function of discrete graphic al models, which can in turn be used, for instance, for marginal computation or model selection.
****
Sparse coding for multitask and transfer learning
Andreas Maurer, Massi Pontil, Bernardino Romera-Paredes
We investigate the use of sparse coding and dictionary learning in the context o f multitask and transfer learning. The central assumption of our learning method is that the tasks parameters are well approximated by sparse linear combination s of the atoms of a dictionary on a high or infinite dimensional space. This ass umption, together with the large quantity of available data in the multitask and transfer learning settings, allows a principled choice of the dictionary. We pr ovide bounds on the generalization error of this approach, for both settings. Nu merical experiments on one synthetic and two real datasets show the advantage of our method over single task learning, a previous method based on orthogonal and dense representation of the tasks and a related method learning task grouping.
****
Direct Modeling of Complex Invariances for Visual Object Features
Ka Yu Hui
View-invariant object representations created from feature pooling networks have been widely adopted in state-of-the-art visual recognition systems. Recently, t he research community seeks to improve these view-invariant representations furt her by additional invariance and receptive field learning, or by taking on the c hallenge of processing massive amounts of learning data. In this paper we consid er an alternate strategy of directly modeling complex invariances of object feat ures. While this may sound like a naive and inferior approach, our experiments s how that this approach can achieve competitive and state-of-the-art accuracy on visual recognition data sets such as CIFAR-10 and STL-10. We present an highly a pplicable dictionary learning algorithm on complex invariances that can be used in most feature pooling network settings. It also has the merits of simplicity a nd requires no additional tuning. We also discuss the implication of our experim ent results concerning recent observations on the usefulness of pre-trained feat ures, and the role of direct invariance modeling in invariance learning.
****
Hierarchically-coupled hidden Markov models for learning kinetic rates from sing le-molecule data
Jan-Willem Meent, Jonathan Bronson, Frank Wood, Ruben Gonzalez Jr., Chris Wiggin s
We address the problem of analyzing sets of noisy time-varying signals that all report on the same process but confound straightforward analyses due to complex inter-signal heterogeneities and measurement artifacts. In particular we consid er single-molecule experiments which indirectly measure the distinct steps in a biomolecular process via observations of noisy time-dependent signals such as a fluorescence intensity or bead position. Straightforward hidden Markov model (HM M) analyses attempt to characterize such processes in terms of a set of conforma tional states, the transitions that can occur between these states, and the asso

ciated rates at which those transitions occur; but require ad-hoc post-processin
g steps to combine multiple signals. Here we develop a hierarchically coupled H
MM that allows experimentalists to deal with inter-signal variability in a princ
ipled and automatic way. Our approach is a generalized expectation maximization
hyperparameter point estimation procedure with variational Bayes at the level of
 individual time series that learns an single interpretable representation of th
e overall data generating process.
****
Activized Learning with Uniform Classification Noise
Liu Yang, Steve Hanneke
We prove that for any VC class, it is possible to transform any passive learning
 algorithm into an active learning algorithm with strong asymptotic improvements
 in label complexity for every nontrivial distribution satisfying a uniform cla
ssification noise condition. This generalizes a similar result proven by (Hann
eke, 2009;2012) for the realizable case, and is the first result establishing t
hat such general improvement guarantees are possible in the presence of restric
ted types of classification noise.
****
Guided Policy Search
Sergey Levine, Vladlen Koltun
Direct policy search can effectively scale to high-dimensional systems, but comp
lex policies with hundreds of parameters often present a challenge for such meth
ods, requiring numerous samples and often falling into poor local optima. We pre
sent a guided policy search algorithm that uses trajectory optimization to direc
t policy learning and avoid poor local optima. We show how differential dynamic
programming can be used to generate suitable guiding samples, and describe a reg
ularized importance sampled policy optimization that incorporates these samples
into the policy search. We evaluate the method by learning neural network contro
llers for planar swimming, hopping, and walking, as well as simulated 3D humanoi
d running.
****
Squared-loss Mutual Information Regularization: A Novel Information-theoretic Ap
proach to Semi-supervised Learning
Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, Masashi Sugiyama
We propose squared-loss mutual information regularization (SMIR) for multi-class
 probabilistic classification, following the information maximization principle.
 SMIR is convex under mild conditions and thus improves the nonconvexity of mutu
al information regularization. It offers all of the following four abilities to
semi-supervised algorithms: Analytical solution, out-of-sample/multi-class class
ification, and probabilistic output. Furthermore, novel generalization error bou
nds are derived. Experiments show SMIR compares favorably with state-of-the-art
methods.
****
Gossip-based distributed stochastic bandit algorithms
Balazs Szorenyi, Robert Busa-Fekete, Istvan Hegedus, Robert Ormandi, Mark Jelasi
ty, Balazs Kegl
The multi-armed bandit problem has attracted remarkable attention in the machine
 learning community and many efficient algorithms have been proposed to handle t
he so-called exploitation-exploration dilemma in various bandit setups. At the s
ame time, significantly less effort has been devoted to adapting bandit algorith
ms to particular architectures, such as sensor networks, multi-core machines, or
 peer-to-peer (P2P) environments, which could potentially speed up their converg
ence. Our goal is to adapt stochastic bandit algorithms to P2P networks. In our
 setup, the same set of arms is available in each peer. In every iteration each
peer can pull one arm independently of the other peers, and then some limited co
mmunication is possible with a few random other peers. As our main result, we s
how that our adaptation achieves a linear speedup in terms of the number of peer
s participating in the network. More precisely, we show that the probability of
 playing a suboptimal arm at a peer in iteration t = Ω(\log N) is proportional
 to 1/(Nt) where N denotes the number of peers. The theoretical results are sup

ported by simulation experiments showing that our algorithm scales gracefully wi th the size of network.
****
The Sample-Complexity of General Reinforcement Learning
Tor Lattimore, Marcus Hutter, Peter Sunehag
We study the sample-complexity of reinforcement learning in a general setting wi thout assuming ergodicity or finiteness of the environment. Instead, we define a topology on the space of environments and show that if an environment class is compact with respect to this topology then finite sample-complexity bounds ar e possible and give an algorithm achieving these bounds. We also show the exis tence of environment classes that are non-compact where finite sample-complexit y bounds are not achievable. A lower bound is presented that matches the upper bound except for logarithmic factors.
****
Hierarchical Regularization Cascade for Joint Learning
Alon Zweig, Daphna Weinshall
As the sheer volume of available benchmark datasets increases, the problem of jo int learning of classifiers and knowledge-transfer between classifiers, becomes more and more relevant. We present a hierarchical approach which exploits inform ation sharing among different classification tasks, in multi-task and multi-clas s settings. It engages a top-down iterative method, which begins by posing an op timization problem with an incentive for large scale sharing among all classes. This incentive to share is gradually decreased,until there is no sharing and all tasks are considered separately. The method therefore exploits different levels of sharing within a given group of related tasks, without having to make hard d ecisions about the grouping of tasks. In order to deal with large scale problems , with many tasks and many classes, we extend our batch approach to an online se tting and provide regret analysis of the algorithm. We tested our approach exten sively on synthetic and real datasets, showing significant improvement over base line and state-of-the-art methods.
****
Multi-Class Classification with Maximum Margin Multiple Kernel
Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh
We present a new algorithm for multi-class classification with multiple kernels. Our algorithm is based on a natural notion of the multi-class margin of a kerne l. We show that larger values of this quantity guarantee the existence of an acc urate multi-class predictor and also define a family of multiple kernel algorith ms based on the maximization of the multi-class margin of a kernel (M^3K). We p resent an extensive theoretical analysis in support of our algorithm, including novel multi-class Rademacher complexity margin bounds. Finally, we also report the results of a series of experiments with several data sets, including compari sons where we improve upon the performance of state-of-the-art algorithms both i n binary and multi-class classification with multiple kernels.
****
Bayesian Games for Adversarial Regression Problems
Michael Großhans, Christoph Sawade, Michael Brückner, Tobias Scheffer
We study regression problems in which an adversary can exercise some control ove r the data generation process. Learner and adversary have conflicting but not ne cessarily perfectly antagonistic objectives. We study the case in which the lear ner is not fully informed about the adversary’s objective; instead, any knowledg e of the learner about parameters of the adversary’s goal may be reflected in a Bayesian prior. We model this problem as a Bayesian game, and characterize condi tions under which a unique Bayesian equilibrium point exists. We experimentally compare the Bayesian equilibrium strategy to the Nash equilibrium strategy, the minimax strategy, and regular linear regression.
****
Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourc ing
Xi Chen, Qihang Lin, Dengyong Zhou
In real crowdsourcing applications, each label from a crowd usually comes with

a certain cost. Given a pre- fixed amount of budget, since different tasks have different ambiguities and different workers have different expertises, we want to find an optimal way to allocate the budget among instance-worker pairs such that the overall label quality can be maximized. To address this issue, we start from the simplest setting in which all workers are assumed to be perfect. We formulate the problem as a Bayesian Markov Decision Process (MDP). Using the dynamic programming (DP) algorithm, one can obtain the optimal allocation policy for a given budget. However, DP is computationally intractable. To solve the computational challenge, we propose a novel approximate policy which is called optimistic knowledge gradient. It is practically efficient while theoretically its consistency can be guaranteed. We then extend the MDP framework to deal with inhomogeneous workers and tasks with contextual information available. The experiments on both simulated and real data demonstrate the superiority of our method.
****

Markov Network Estimation From Multi-attribute Data

Mladen Kolar, Han Liu, Eric Xing

Many real world network problems often concern multivariate nodal attributes such as image, textual, and multi-view feature vectors on nodes, rather than simple univariate nodal attributes. The existing graph estimation methods built on Gaussian graphical models and covariance selection algorithms can not handle such data, neither can the theories developed around such methods be directly applied. In this paper, we propose a new principled framework for estimating multi-attribute graphs. Instead of estimating the partial correlation as in current literature, our method estimates the partial canonical correlations that naturally accommodate complex nodal features. Computationally, we provide an efficient algorithm which utilizes the multi-attribute structure. Theoretically, we provide sufficient conditions which guarantee consistent graph recovery. Extensive simulation studies demonstrate performance of our method under various conditions.
****

MILEAGE: Multiple Instance LEArning with Global Embedding

Dan Zhang, Jingrui He, Luo Si, Richard Lawrence

Multiple Instance Learning (MIL) methods generally represent each example as a collection of instances such that the features for local objects can be better captured, whereas traditional learning methods typically extract a global feature vector for each example as an integral part. However, there is limited research work on which of the two learning scenarios performs better. This paper proposes a novel framework – \emphMultiple Instance LEArning with Global Embedding (MILEAGE), in which the global feature vectors for traditional learning methods are integrated into the MIL setting. MILEAGE can leverage the benefits derived from both learning settings. Within the proposed framework, a large margin method is formulated. In particular, the proposed method adaptively tunes the weights on the two different kinds of feature representations (i.e., global and multiple instance) for each example and trains the classifier simultaneously. An alternative algorithm is proposed to solve the resulting optimization problem, which extends the bundle method to the non-convex case. Some important properties of the proposed method, such as the convergence rate and the generalization error rate, are analyzed. A series of experiments have been conducted to demonstrate the advantages of the proposed method over several state-of-the-art multiple instance and traditional learning methods.
****

Guaranteed Sparse Recovery under Linear Transformation

Ji Liu, Lei Yuan, Jieping Ye

We consider the following signal recovery problem: given a measurement matrix $\Phi\in\mathbb{R}^{n\times p}$ and a noisy observation vector $c\in\mathbb{R}^n$ constructed from $c = \Phi\theta^* + \epsilon$ where $\epsilon\in\mathbb{R}^n$ is the noise vector whose entries follow i.i.d. centered sub-Gaussian distribution, how to recover the signal θ^* if $D\theta^*$ is sparse \rca under a linear transformation $D\in\mathbb{R}^{m\times p}$? One natural method using convex optimization is to solve the following problem: $\min_\theta {1\over 2}\|\Phi\theta - c\|^2 + \lambda\|D\theta\|_1$. This paper provides an upper bound of the estimate error and shows the consistency property of this method by assuming that the

design matrix Φis a Gaussian random matrix. Specifically, we show 1) in the noi
seless case, if the condition number of D is bounded and the measurement number
 n≥Ω(s\log(p)) where s is the sparsity number, then the true solution can be re
covered with high probability; and 2) in the noisy case, if the condition numbe
r of D is bounded and the measurement increases faster than s\log(p), that is,
s\log(p)=o(n), the estimate error converges to zero with probability 1 when p a
nd s go to infinity. Our results are consistent with those for the special cas
e D=\boldI_p\times p (equivalently LASSO) and improve the existing analysis. Th
e condition number of D plays a critical role in our analysis. We consider the
condition numbers in two cases including the fused LASSO and the random graph:
the condition number in the fused LASSO case is bounded by a constant, while t
he condition number in the random graph case is bounded with high probability i
f m\over p (i.e., #\textedge\over #\textvertex$) is larger than a certain const
ant. Numerical simulations are consistent with our theoretical results.
****
Learning invariant features by harnessing the aperture problem
Roland Memisevic, Georgios Exarchakis
The energy model is a simple, biologically inspired approach to extracting relat
ionships between images in tasks like stereopsis and motion analysis. We discuss
 how adding an extra pooling layer to the energy model makes it possible to lear
n encodings of transformations that are mostly invariant with respect to image c
ontent, and to learn encodings of images that are mostly invariant with respect
to the observed transformations. We show how this makes it possible to learn 3D
pose-invariant features of objects by watching videos of the objects. We test ou
r approach on a dataset of videos derived from the NORB dataset.
****
Efficient Ranking from Pairwise Comparisons
Fabian Wauthier, Michael Jordan, Nebojsa Jojic
The ranking of n objects based on pairwise comparisons is a core machine learnin
g problem, arising in recommender systems, ad placement, player ranking, biologi
cal applications and others. In many practical situations the true pairwise comp
arisons cannot be actively measured, but a subset of all n(n-1)/2 comparisons is
 passively and noisily observed. Optimization algorithms (e.g., the SVM) could b
e used to predict a ranking with fixed expected Kendall tau distance, while achi
eving an Ω(n) lower bound on the corresponding sample complexity. However, due t
o their centralized structure they are difficult to extend to online or distribu
ted settings. In this paper we show that much simpler algorithms can match the s
ame Ω(n) lower bound in expectation. Furthermore, if an average of O(n\log(n)) b
inary comparisons are measured, then one algorithm recovers the true ranking in
a uniform sense, while the other predicts the ranking more accurately near the t
op than the bottom. We discuss extensions to online and distributed ranking, wit
h benefits over traditional alternatives.
****
Differentially Private Learning with Kernels
Prateek Jain, Abhradeep Thakurta
In this paper, we consider the problem of differentially private learning where
access to the training features is through a kernel function only. Existing work
 on this problem is restricted to translation invariant kernels only, where (app
roximate) training features are available explicitly. In fact, for general clas
s of kernel functions and in general setting of releasing different private pred
ictor (\w^*), the problem is impossible to solve \citeCMS11. In this work, we re
lax the problem setting into three different easier but practical settings. In o
ur first problem setting, we consider an interactive model where the user sends
its test set to a trusted learner who sends back differentially private predicti
ons over the test points. This setting is prevalent in modern online systems lik
e search engines, ad engines etc. In the second model, the learner sends back a
differentially private version of the optimal parameter vector \w^* but requires
 access to a small subset of unlabeled test set beforehand. This also is a pract
ical setting that involves two parties interacting through trusted third party.
Our third model is similar to the traditional model, where learner is oblivious

to the test set and needs to send a differentially private version of \w^*, but the kernels are restricted to efficiently computable functions over low-dimensional vector spaces. For each of the models, we derive differentially private learning algorithms with provable “utlity” or error bounds. Moreover, we show that our methods can also be applied to the traditional setting of \cite Rubinstein09, CMS11. Here, our sample complexity bounds have only O(d^1/3) dependence on the dimensionality d while existing methods require O(d^1/2) samples to achieve same generalization error.
****
Thompson Sampling for Contextual Bandits with Linear Payoffs
Shipra Agrawal, Navin Goyal
Thompson Sampling is one of the oldest heuristics for multi-armed bandit problems. It is a randomized algorithm based on Bayesian ideas, and has recently generated significant interest after several studies demonstrated it to have better empirical performance compared to the state of the art methods. However, many questions regarding its theoretical performance remained open. In this paper, we design and analyze Thompson Sampling algorithm for the stochastic contextual multi-armed bandit problem with linear payoff functions, when the contexts are provided by an adaptive adversary. This is among the most important and widely studied version of the contextual bandits problem. We prove a high probability regret bound of \tildeO(\fracd\sqrtε\sqrtT^1+ε) in time T for any ε∈(0,1), where d is the dimension of each context vector and εis a parameter used by the algorithm. Our results provide the first theoretical guarantees for the contextual version of Thompson Sampling, and are close to the lower bound of Ω(\sqrtdT) for this problem. This essentially solves the COLT open problem of Chapelle and Li [COLT 2012] regarding regret bounds for Thompson Sampling for contextual bandits problem with linear payoff functions.   Our version of Thompson sampling uses Gaussian prior and Gaussian likelihood function. Our novel martingale-based analysis techniques also allow easy extensions to the use of other distributions, satisfying certain general conditions.
****
Learning Multiple Behaviors from Unlabeled Demonstrations in a Latent Controller Space
Javier Almingol, Lui Montesano, Manuel Lopes
In this paper we introduce a method to learn multiple behaviors in the form of motor primitives from an unlabeled dataset. One of the difficulties of this problem is that in the measurement space, behaviors can be very mixed, despite existing a latent representation where they can be easily separated. We propose a mixture model based on Dirichlet Process (DP) to simultaneously cluster the observed time-series and recover a sparse representation of the behaviors using a Laplacian prior as the base measure of the DP. We show that for linear models, e.g potential functions generated by linear combinations of a large number of features, it is possible to compute analytically the marginal of the observations and derive an efficient sampler. The method is evaluated using robot behaviors and real data from human motion and compared to other techniques.
****
Inference algorithms for pattern-based CRFs on sequence data
Rustem Takhanov, Vladimir Kolmogorov
We consider \em Conditional Random Fields (CRFs) with pattern-based potentials defined on a chain. In this model the energy of a string (labeling) x_1\ldots x_n is the sum of terms over intervals [i,j] where each term is non-zero only if the substring x_i\ldots x_j equals a prespecified pattern α. Such CRFs can be naturally applied to many sequence tagging problems. We present efficient algorithms for the three standard inference tasks in a CRF, namely computing (i) the partition function, (ii) marginals, and (iii) computing the MAP. Their complexities are respectively O(n L), O(n L \ell_\max) and O(n L \min{|D|,\log (\ell_\max + 1)}) where L is the combined length of input patterns, \ell_\max is the maximum length of a pattern, and D is the input alphabet. This improves on the previous algorithms of \citeYe:NIPS09 whose complexities are respectively O(n L |D|), O\left(n |Γ| L^2 \ell_\max^2\right) and O(n L |D|), where |Γ| is the numb

er of input patterns. In addition, we give an efficient algorithm for sampling,
 and revisit the case of MAP with non-positive weights. Finally, we apply patter
n-based CRFs to the problem of the protein dihedral angles prediction.
****
One-Bit Compressed Sensing: Provable Support and Vector Recovery
Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, Aditya Nori
In this paper, we study the problem of one-bit compressed sensing (1-bit CS), wh
ere the goal is to design a measurement matrix A and a recovery algorithm s.t. a
 k-sparse vector \x^* can be efficiently recovered back from signed linear measu
rements, i.e., b=\sign(A\x^*). This is an important problem in the signal acquis
ition area and has several learning applications as well, e.g., multi-label clas
sification \citeHsuKLZ10. We study this problem in two settings: a) support reco
very: recover \supp(\x^*), b) approximate vector recovery: recover a unit vector
 \hx s.t. || \hatx-\x^*/||\x^*|| ||_2≤ε. For support recovery, we propose two no
vel and efficient solutions based on two combinatorial structures: union free fa
mily of sets and expanders. In contrast to existing methods for support recove
ry, our methods are universal i.e. a single measurement matrix A can recover alm
ost all the signals. For approximate recovery, we propose the first method to r
ecover sparse vector using a near optimal number of measurements. We also empir
ically demonstrate effectiveness of our algorithms; we show that our algorithms
 are able to recover signals with smaller number of measurements than several ex
isting methods.
****
Tensor Analyzers
Yichuan Tang, Ruslan Salakhutdinov, Geoffrey Hinton
Factor Analysis is a statistical method that seeks to explain linear variations
in data by using unobserved latent variables. Due to its additive nature, it is
not suitable for modeling data that is generated by multiple groups of latent fa
ctors which interact multiplicatively. In this paper, we introduce Tensor Analyz
ers which are a multilinear generalization of Factor Analyzers. We describe an e
fficient way of sampling from the posterior distribution over factor values and
we demonstrate that these samples can be used in the EM algorithm for learning i
nteresting mixture models of natural image patches. Tensor Analyzers can also ac
curately recognize a face under significant pose and illumination variations whe
n given only one previous image of that face. We also show that Tensor Analyzers
 can be trained in an unsupervised, semi-supervised, or fully supervised setting
s.
****
Learning Sparse Penalties for Change-point Detection using Max Margin Interval R
egression
Toby Hocking, Guillem Rigaill, Jean-Philippe Vert, Francis Bach
In segmentation models, the number of change-points is typically   chosen using
 a penalized cost function. In this work, we propose to   learn the penalty an
d its constants in databases of signals with   weak change-point annotations. W
e propose a convex relaxation for   the resulting interval regression problem,
and solve it using   accelerated proximal gradient methods. We show that this m
ethod   achieves state-of-the-art change-point detection in a database of   an
notated DNA copy number profiles from neuroblastoma tumors.
****
Learning from Human-Generated Lists
Kwang-Sung Jun, Jerry Zhu, Burr Settles, Timothy Rogers
Human-generated lists are a form of non-iid data with important applications in
machine learning and cognitive psychology. We propose a generative model - sampl
ing with reduced replacement (SWIRL) - for such lists. We discuss SWIRL’s relati
on to standard sampling paradigms, provide the maximum likelihood estimate for l
earning, and demonstrate its value with two real-world applications: (i) In a ""
feature volunteering"" task where non-experts spontaneously generate feature=>la
bel pairs for text classification, SWIRL improves the accuracy of state-of-the-a
rt feature-learning frameworks. (ii) In a ""verbal fluency"" task where brain-da
maged patients generate word lists when prompted with a category, SWIRL paramete

rs align well with existing psychological theories, and our model can classify healthy people vs. patients from the lists they generate.
****
A Fast and Exact Energy Minimization Algorithm for Cycle MRFs
Huayan Wang, Koller Daphne
The presence of cycles gives rise to the difficulty in performing inference for MRFs. Handling cycles efficiently would greatly enhance our ability to tackle general MRFs. In particular, for dual decomposition of energy minimization (MAP inference), using cycle subproblems leads to a much tighter relaxation than using trees, but solving the cycle subproblems turns out to be the bottleneck. In this paper, we present a fast and exact algorithm for energy minimization in cycle MRFs, which can be used as a subroutine in tackling general MRFs. Our method builds on junction-tree message passing, with a large portion of the message entries pruned for efficiency. The pruning conditions fully exploit the structure of a cycle. Experimental results show that our algorithm is more than an order of magnitude faster than other state-of-the-art fast inference methods, and it performs consistently well in several different real problems.
****
Stochastic k-Neighborhood Selection for Supervised and Unsupervised Learning
Daniel Tarlow, Kevin Swersky, Laurent Charlin, Ilya Sutskever, Rich Zemel
Neighborhood Components Analysis (NCA) is a popular method for learning a distance metric to be used within a k-nearest neighbors (kNN) classifier. A key assumption built into the model is that each point stochastically selects a single neighbor, which makes the model well-justified only for kNN with k=1. However, kNN classifiers with k>1 are more robust and usually preferred in practice. Here we present kNCA, which generalizes NCA by learning distance metrics that are appropriate for kNN with arbitrary k. The main technical contribution is showing how to efficiently compute and optimize the expected accuracy of a kNN classifier. We apply similar ideas in an unsupervised setting to yield kSNE and ktSNE, generalizations of Stochastic Neighbor Embedding (SNE, tSNE) that operate on neighborhoods of size k, which provide an axis of control over embeddings that allow for more homogeneous and interpretable regions. Empirically, we show that kNCA often improves classification accuracy over state of the art methods, produces qualitative differences in the embeddings as k is varied, and is more robust with respect to label noise.
****
An Efficient Posterior Regularized Latent Variable Model for Interactive Sound Source Separation
Nicholas Bryan, Gautham Mysore
In applications such as audio denoising, music transcription, music remixing, and audio-based forensics, it is desirable to decompose a single-channel recording into its respective sources. One of the current most effective class of methods to do so is based on non-negative matrix factorization and related latent variable models. Such techniques, however, typically perform poorly when no isolated training data is given and do not allow user feedback to correct for poor results. To overcome these issues, we allow a user to interactively constrain a latent variable model by painting on a time-frequency display of sound to guide the learning process. The annotations are used within the framework of posterior regularization to impose linear grouping constraints that would otherwise be difficult to achieve via standard priors. For the constraints considered, an efficient expectation-maximization algorithm is derived with closed-form multiplicative updates, drawing connections to non-negative matrix factorization methods, and allowing for high-quality interactive-rate separation without explicit training data.
****
Estimating Unknown Sparsity in Compressed Sensing
Miles Lopes
In the theory of compressed sensing (CS), the sparsity $\|x\|_0$ of the unknown signal $x\in\R^p$ is commonly assumed to be a known parameter. However, it is typically unknown in practice. Due to the fact that many aspects of CS depend on knowin

g $\|x\|_0$, it is important to estimate this parameter in a data-driven way. A second practical concern is that $\|x\|_0$ is a highly unstable function of x. In particular, for real signals with entries not exactly equal to 0, the value $\|x\|_0=p$ is not a useful description of the effective number of coordinates. In this paper, we propose to estimate a stable measure of sparsity $s(x):=\|x\|_1^2/\|x\|_2^2$, which is a sharp lower bound on $\|x\|_0$. Our estimation procedure uses only a small number of linear measurements, does not rely on any sparsity assumptions, and requires very little computation. A confidence interval for s(x) is provided, and its width is shown to have no dependence on the signal dimension p. Moreover, this result extends naturally to the matrix recovery setting, where a soft version of matrix rank can be estimated with analogous guarantees. Finally, we show that the use of randomized measurements is essential to estimating s(x). This is accomplished by proving that the minimax risk for estimating s(x) with deterministic measurements is large when n■p.
****

MAD-Bayes: MAP-based Asymptotic Derivations from Bayes

Tamara Broderick, Brian Kulis, Michael Jordan

The classical mixture of Gaussians model is related to K-means via small-variance asymptotics: as the covariances of the Gaussians tend to zero, the negative log-likelihood of the mixture of Gaussians model approaches the K-means objective, and the EM algorithm approaches the K-means algorithm. Kulis & Jordan (2012) used this observation to obtain a novel K-means-like algorithm from a Gibbs sampler for the Dirichlet process (DP) mixture. We instead consider applying small-variance asymptotics directly to the posterior in Bayesian nonparametric models. This framework is independent of any specific Bayesian inference algorithm, and it has the major advantage that it generalizes immediately to a range of models beyond the DP mixture. To illustrate, we apply our framework to the feature learning setting, where the beta process and Indian buffet process provide an appropriate Bayesian nonparametric prior. We obtain a novel objective function that goes beyond clustering to learn (and penalize new) groupings for which we relax the mutual exclusivity and exhaustivity assumptions of clustering. We demonstrate several other algorithms, all of which are scalable and simple to implement. Empirical results demonstrate the benefits of the new framework.
****

The Most Generative Maximum Margin Bayesian Networks

Robert Peharz, Sebastian Tschiatschek, Franz Pernkopf

Although discriminative learning in graphical models generally improves classification results, the generative semantics of the model are compromised. In this paper, we introduce a novel approach of hybrid generative-discriminative learning for Bayesian networks. We use an SVM-type large margin formulation for discriminative training, introducing a likelihood-weighted ℓ^1-norm for the SVM-norm-penalization. This simultaneously optimizes the data likelihood and therefore partly maintains the generative character of the model. For many network structures, our method can be formulated as a convex problem, guaranteeing a globally optimal solution. In terms of classification, the resulting models outperform state-of-the art generative and discriminative learning methods for Bayesian networks, and are comparable with linear and kernelized SVMs. Furthermore, the models achieve likelihoods close to the maximum likelihood solution and show robust behavior in classification experiments with missing features.
****

Fastfood - Computing Hilbert Space Expansions in loglinear time

Quoc Le, Tamas Sarlos, Alexander Smola

Fast nonlinear function classes are crucial for nonparametric estimation, such as in kernel methods. This paper proposes an improvement to random kitchen sinks that offers significantly faster computation in log-linear time without sacrificing accuracy. Furthermore, we show how one may adjust the regularization properties of the kernel simply by changing the spectral distribution of the projection matrix. We provide experimental results which show that even for for moderately small problems we already achieve two orders of magnitude faster computation and three orders of magnitude lower memory footprint.

****
Joint Transfer and Batch-mode Active Learning
Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, Jieping Ye
Active learning and transfer learning are two different methodologies that address the common problem of insufficient labels. Transfer learning addresses this problem by using the knowledge gained from a related and already labeled data source, whereas active learning focuses on selecting a small set of informative samples for manual annotation. Recently, there has been much interest in developing frameworks that combine both transfer and active learning methodologies. A few such frameworks reported in literature perform transfer and active learning in two separate stages. In this work, we present an integrated framework that performs transfer and active learning simultaneously by solving a single convex optimization problem. The proposed framework computes the weights of source domain data and selects the samples from the target domain data simultaneously, by minimizing a common objective of reducing distribution difference between the data set consisting of reweighted source and the queried target domain data and the set of unlabeled target domain data. Comprehensive experiments on three real world data sets demonstrate that the proposed method improves the classification accuracy by 5% to 10% over the existing two-stage approach
****
Message passing with l1 penalized KL minimization
Yuan Qi, Yandong Guo
Bayesian inference is often hampered by large computational expense. As a generalization of belief propagation (BP), expectation propagation (EP) approximates exact Bayesian computation with efficient message passing updates. However, when an approximation family used by EP is far from exact posterior distributions, message passing may lead to poor approximation quality and suffer from divergence. To address this issue, we propose an approximate inference method, relaxed expectation propagation(REP), based on a new divergence with a l1 penalty. Minimizing this penalized divergence adaptively relaxes EP's moment matching requirement for message passing. We apply REP to Gaussian process classification and experimental results demonstrate significant improvement of REP over EP and alpha-divergence based power EP – in terms of algorithmic stability, estimation accuracy, and predictive performance. Furthermore, we develop relaxed belief propagation(RBP), a special case of REP, to conduct inference on discrete Markov random fields (MRFs). Our results show improved estimation accuracy of RBP over BP and fractional BP when interactions between MRF nodes are strong.
****
Mean Reversion with a Variance Threshold
Marco Cuturi, Alexandre D'Aspremont
Starting from a multivariate data set, we study several techniques to isolate affine combinations of the variables with a maximum amount of mean reversion, while constraining the variance to be larger than a given threshold. We show that many of the optimization problems arising in this context can be solved exactly using semidefinite programming and some variant of the \mathcalS-lemma. In finance, these methods are used to isolate statistical arbitrage opportunities, i.e. mean reverting portfolios with enough variance to overcome market friction. In a more general setting, mean reversion and its generalizations are also used as a proxy for stationarity, while variance simply measures signal strength.
****
Top-down particle filtering for Bayesian decision trees
Balaji Lakshminarayanan, Daniel Roy, Yee Whye Teh
Decision tree learning is a popular approach for classification and regression in machine learning and statistics, and Bayesian formulations - which introduce a prior distribution over decision trees, and formulate learning as posterior inference given data - have been shown to produce competitive performance. Unlike classic decision tree learning algorithms like ID3, C4.5 and CART, which work in a top-down manner, existing Bayesian algorithms produce an approximation to the posterior distribution by evolving a complete tree (or collection thereof) iteratively via local Monte Carlo modifications to the structure of the tree, e.g., u

sing Markov chain Monte Carlo (MCMC). We present a sequential Monte Carlo (SMC) algorithm that instead works in a top-down manner, mimicking the behavior and speed of classic algorithms. We demonstrate empirically that our approach delivers accuracy comparable to the most popular MCMC method, but operates more than an order of magnitude faster, and thus represents a better computation-accuracy tradeoff.
****
Smooth Sparse Coding via Marginal Regression for Learning Sparse Representations
Krishnakumar Balasubramanian, Kai Yu, Guy Lebanon
We propose and analyze a novel framework for learning sparse representations, based on two statistical techniques: kernel smoothing and marginal regression. The proposed approach provides a flexible framework for incorporating feature similarity or temporal information present in data sets, via nonparametric kernel smoothing. We provide generalization bounds for dictionary learning using smooth sparse coding and show how the sample complexity depends on the L1 norm of kernel function used. Furthermore, we propose using marginal regression for obtaining sparse codes, which significantly improves the speed and allows one to scale to large dictionary sizes easily. We demonstrate the advantages of the proposed approach, both in terms of accuracy and speed by extensive experimentation on several real data sets. In addition, we demonstrate how the proposed approach could be used for improving semisupervised sparse coding.
****
Robust and Discriminative Self-Taught Learning
Hua Wang, Feiping Nie, Heng Huang
The lack of training data is a common challenge in many machine learning problems, which is often tackled by semi-supervised learning methods or transfer learning methods. The former requires unlabeled images from the same distribution as the labeled ones and the latter leverages labeled images from related homogenous tasks. However, these restrictions often cannot be satisfied. To address this, we propose a novel robust and discriminative self-taught learning approach to utilize any unlabeled data without the above restrictions. Our new approach employs a robust loss function to learn the dictionary, and enforces the structured sparse regularization to automatically select the optimal dictionary basis vectors and incorporate the supervision information contained in the labeled data. We derive an efficient iterative algorithm to solve the optimization problem and rigorously prove its convergence. Promising results in extensive experiments have validated the proposed approach.
****
Safe Policy Iteration
Matteo Pirotta, Marcello Restelli, Alessio Pecorino, Daniele Calandriello
This paper presents a study of the policy improvement step that can be usefully exploited by approximate policy-iteration algorithms. When either the policy evaluation step or the policy improvement step returns an approximated result, the sequence of policies produced by policy iteration may not be monotonically increasing, and oscillations may occur. To address this issue, we consider safe policy improvements, i.e., at each iteration we search for a policy that maximizes a lower bound to the policy improvement w.r.t. the current policy. When no improving policy can be found the algorithm stops. We propose two safe policy-iteration algorithms that differ in the way the next policy is chosen w.r.t. the estimated greedy policy. Besides being theoretically derived and discussed, the proposed algorithms are empirically evaluated and compared with state-of-the-art approaches on some chain-walk domains and on the Blackjack card game.
****
Unfolding Latent Tree Structures using 4th Order Tensors
Mariya Ishteva, Haesun Park, Le Song
Discovering the latent structure from many observed variables is an important yet challenging learning task. Existing approaches for discovering latent structures often require the unknown number of hidden states as an input. In this paper, we propose a quartet based approach which is agnostic to this number. The key contribution is a novel rank characterization of the tensor associated with the m

arginal distribution of a quartet. This characterization allows us to design a n
uclear norm based test for resolving quartet relations. We then use the quartet
test as a subroutine in a divide-and-conquer algorithm for recovering the latent
 tree structure. Under mild conditions, the algorithm is consistent and its erro
r probability decays exponentially with increasing sample size. We demonstrate t
hat the proposed approach compares favorably to alternatives. In a real world st
ock dataset, it also discovers meaningful groupings of variables, and produces a
 model that fits the data better.
****
Learning Fair Representations
Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork
We propose a learning algorithm for fair classification that achieves both group
 fairness (the proportion of members in a protected group receiving positive cla
ssification is identical to the proportion in the population as a whole), and i
ndividual fairness (similar individuals should be treated similarly). We formul
ate fairness as an optimization problem of finding a good representation of the
 data with two competing goals: to encode the data as well as possible, while si
multaneously obfuscating any information about membership in the protected group
. We show positive results of our algorithm relative to other known techniques,
 on three datasets. Moreover, we demonstrate several advantages to our approach
. First, our intermediate representation can be used for other classification t
asks (i.e., transfer learning is possible); secondly, we take a step toward lea
rning a distance metric which can find important dimensions of the data for clas
sification.
****
Hierarchical Tensor Decomposition of Latent Tree Graphical Models
Le Song, Mariya Ishteva, Ankur Parikh, Eric Xing, Haesun Park
We approach the problem of estimating the parameters of a latent tree graphical
model from a hierarchical tensor decomposition point of view. In this new view,
the marginal probability table of the observed variables in a latent tree is tre
ated as a tensor, and we show that: (i) the latent variables induce low rank str
uctures in various matricizations of the tensor; (ii) this collection of low ran
k matricizations induce a hierarchical low rank decomposition of the tensor. Exp
loiting these properties, we derive an optimization problem for estimating the p
arameters of a latent tree graphical model, i.e., hierarchical decomposion of a
tensor which minimizes the Frobenius norm of the difference between the original
 tensor and its decomposition.   When the latent tree graphical models are corr
ectly specified, we show that a global optimum of the optimization problem can b
e obtained via a recursive decomposition algorithm. This algorithm recovers prev
ious spectral algorithms for hidden Markov models (Hsu et al., 2009; Foster et a
l., 2012) and latent tree graphical models (Parikh et al., 2011; Song et al., 20
11) as special cases, elucidating the global objective these algorithms are opti
mizing for. When the latent tree graphical models are misspecified, we derive a
better decomposition based on our framework, and provide approximation guarantee
 for this new estimator. In both synthetic and real world data, this new estimat
or significantly improves over the-state-of-the-art.
****
No more pesky learning rates
Tom Schaul, Sixin Zhang, Yann LeCun
The performance of stochastic gradient descent (SGD) depends critically on how l
earning rates are tuned and decreased over time. We propose a method to automati
cally adjust multiple learning rates so as to minimize the expected error at any
 one time. The method relies on local gradient variations across samples. In our
 approach, learning rates can increase as well as decrease, making it suitable f
or non-stationary problems. Using a number of convex and non-convex learning tas
ks, we show that the resulting algorithm matches the performance of the best set
tings obtained through systematic search, and effectively removes the need for l
earning rate tuning.
****
Multi-View Clustering and Feature Learning via Structured Sparsity

Hua Wang, Feiping Nie, Heng Huang
Combining information from various data sources has become an important research topic in machine learning with many scientific applications. Most previous studies employ kernels or graphs to integrate different types of features, which routinely assume one weight for one type of features. However, for many problems, the importance of features in one source to an individual cluster of data can be varied, which makes the previous approaches ineffective. In this paper, we propose a novel multi-view learning model to integrate all features and learn the weight for every feature with respect to each cluster individually via new joint structured sparsity-inducing norms. The proposed multi-view learning framework allows us not only to perform clustering tasks, but also to deal with classification tasks by an extension when the labeling knowledge is available. A new efficient algorithm is derived to solve the formulated objective with rigorous theoretical proof on its convergence. We applied our new data fusion method to five broadly used multi-view data sets for both clustering and classification. In all experimental results, our method clearly outperforms other related state-of-the-art methods.
****
Planning by Prioritized Sweeping with Small Backups
Harm Van Seijen, Rich Sutton
Efficient planning plays a crucial role in model-based reinforcement learning. Traditionally, the main planning operation is a full backup based on the current estimates of the successor states. Consequently, its computation time is proportional to the number of successor states. In this paper, we introduce a new planning backup that uses only the current value of a single successor state and has a computation time independent of the number of successor states. This new backup, which we call a small backup, opens the door to a new class of model-based reinforcement learning methods that exhibit much finer control over their planning process than traditional methods. We empirically demonstrate that this increased flexibility allows for more efficient planning by showing that an implementation of prioritized sweeping based on small backups achieves a substantial performance improvement over classical implementations.
****
Solving Continuous POMDPs: Value Iteration with Incremental Learning of an Efficient Space Representation
Sebastian Brechtel, Tobias Gindele, Rüdiger Dillmann
Discrete POMDPs of medium complexity can be approximately solved in reasonable time. However, most applications have a continuous and thus uncountably infinite state space. We propose the novel concept of learning a discrete representation of the continuous state space to solve the integrals in continuous POMDPs efficiently and generalize sparse calculations over the continuous space. The representation is iteratively refined as part of a novel Value Iteration step and does not depend on prior knowledge. Consistency for the learned generalization is asserted by a self-correction algorithm. The presented concept is implemented for continuous state and observation spaces based on Monte Carlo approximation to allow for arbitrary POMDP models. In an experimental comparison it yields higher values in significantly shorter time than state of the art algorithms and solves higher-dimensional problems.
****
Learning Heteroscedastic Models by Convex Programming under Group Sparsity
Arnak Dalalyan, Mohamed Hebiri, Katia Meziani, Joseph Salmon
Sparse estimation methods based on l1 relaxation, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a   major obstacle in applying these methods in several frameworks, such as time series, random fields, inverse problems, for which noise is rarely homoscedastic and the noise level is hard to know in advance. In this paper, we propose a new approach to the joint estimation of the conditional mean and the conditional variance in a high-dimensional (auto-) regression setting.   An attractive feature of the proposed estimator is   that it is efficiently computable even for very large   s

cale problems by solving a second-order cone program (SOCP). We present theoretical analysis and numerical results assessing the performance of the proposed procedure.
****
Covariate Shift in Hilbert Space: A Solution via Sorrogate Kernels
Kai Zhang, Vincent Zheng, Qiaojun Wang, James Kwok, Qiang Yang, Ivan Marsic
Covariate shift is a unconventional learning scenario in which training and testing data have different distributions. A general principle to solve the problem is to make the training data distribution similar to the test one, such that classifiers computed on the former generalizes well to the latter. Current approaches typically target on the sample distribution in the input space, however, for kernel-based learning methods, the algorithm performance depends directly on the geometry of the kernel-induced feature space. Motivated by this, we propose to match data distributions in the Hilbert space, which, given a pre-defined empirical kernel map, can be formulated as aligning kernel matrices across domains. In particular, to evaluate similarity of kernel matrices defined on arbitrarily different samples, the novel concept of surrogate kernel is introduced based on the Mercer's theorem. Our approach caters the model adaptation specifically to kernel-based learning mechanism, and demonstrates promising results on several real-world applications.
****
A Local Algorithm for Finding Well-Connected Clusters
Zeyuan Allen Zhu, Silvio Lattanzi, Vahab Mirrokni
Motivated by applications of large-scale graph clustering, we study random-walk-based LOCAL algorithms whose running times depend only on the size of the output cluster, rather than the entire graph. In particular, we develop a method with better theoretical guarantee compared to all previous work, both in terms of the clustering accuracy and the conductance of the output set. We also prove that our analysis is tight, and perform empirical evaluation to support our theory on both synthetic and real data. More specifically, our method outperforms prior work when the cluster is WELL-CONNECTED. In fact, the better it is well-connected inside, the more significant improvement we can obtain. Our results shed light on why in practice some random-walk-based algorithms perform better than its previous theory, and help guide future research about local clustering.
****
Efficient Multi-label Classification with Many Labels
Wei Bi, James Kwok
Multi-label classification deals with the problem where each instance can be associated with a set of class labels. However, in many real-world applications, the number of class labels can be in the hundreds or even thousands, and existing multi-label classification methods often become computationally inefficient. In recent years, a number of remedies have been proposed. However, they are either based on simple dimension reduction techniques or involve expensive optimization problems. In this paper, we address this problem by selecting a small subset of class labels that can approximately span the original label space. This is performed by randomized sampling where the sampling probability of each class label reflects its importance among all the labels. Theoretical analysis shows that this randomized sampling approach is highly efficient. Experiments on a number of real-world multi-label datasets with many labels demonstrate the appealing performance and efficiency of the proposed algorithm.
****
Spectral Compressed Sensing via Structured Matrix Completion
Yuxin Chen, Yuejie Chi
The paper studies the problem of recovering a spectrally sparse object from a small number of time domain samples. Specifically, the object of interest with ambient dimension n is assumed to be a mixture of r complex multi-dimensional sinusoids, while the underlying frequencies can assume any value in the unit disk. Conventional compressed sensing paradigms suffer from the \em basis mismatch issue when imposing a discrete dictionary on the Fourier representation. To address this problem, we develop a novel nonparametric algorithm, called enhanced matrix

completion (EMaC), based on structured matrix completion. The algorithm starts by converting the data into a low-rank enhanced form with multi-fold Hankel stru cture, then attempts recovery via nuclear norm minimization. Under mild incohere nce conditions, EMaC allows perfect recovery as soon as the number of samples ex ceeds the order of \mathcalO(r\log^2 n). We also show that, in many instances, a ccurate completion of a low-rank multi-fold Hankel matrix is possible when the n umber of observed entries is proportional to the information theoretical limits (except for a logarithmic gap). The robustness of EMaC against bounded noise and its applicability to super resolution are further demonstrated by numerical exp eriments.
****
Multi-Task Learning with Gaussian Matrix Generalized Inverse Gaussian Model
Ming Yang, Yingming Li, Zhongfei Zhang
In this paper, we study the multi-task learning problem with a new perspective o f considering the structure of the residue error matrix and the low-rank approxi mation to the task covariance matrix simultaneously. In particular, we first int roduce the Matrix Generalized Inverse Gaussian (MGIG) prior and define a Gaussia n Matrix Generalized Inverse Gaussian (GMGIG) model for low-rank approximation t o the task covariance matrix. Through combining the GMGIG model with the residua l error structure assumption, we propose the GMGIG regression model for multi-ta sk learning. To make the computation tractable, we simultaneously use variationa l inference and sampling techniques. In particular, we propose two sampling stra tegies for computing the statistics of the MGIG distribution. Experiments show t hat this model is superior to the peer methods in regression and prediction.
****
Simple Sparsification Improves Sparse Denoising Autoencoders in Denoising Highly Corrupted Images
Kyunghyun Cho
Recently Burger et al. (2012) and Xie et al. (2012) proposed to use a denoising autoencoder (DAE) for denoising noisy images. They showed that a plain, deep DAE can denoise noisy images as well as the conventional methods such as BM3D and K SVD. Both of them approached image denoising by denoising small, image patches o f a larger image and combining them to form a clean image. In this setting, it i s usual to use the encoder of the DAE to obtain the latent representation and su bsequently apply the decoder to get the clean patch. We propose that a simple sp arsification of the latent representation found by the encoder improves denoisin g performance, when the DAE was trained with sparsity regularization. The experi ments confirm that the proposed sparsification indeed helps both denoising a sma ll image patch and denoising a larger image consisting of those patches. Further more, it is found out that the proposed method improves even classification perf ormance when test samples are corrupted with noise.
****
On the Generalization Ability of Online Learning Algorithms for Pairwise Loss Fu nctions
Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, Harish Karnick
In this paper, we study the generalization properties of online learning based s tochastic methods for supervised learning problems where the loss function is de pendent on more than one training sample (e.g., metric learning, ranking). We pr esent a generic decoupling technique that enables us to provide Rademacher compl exity-based generalization error bounds. Our bounds are in general tighter than those obtained by Wang et al. (COLT 2012) for the same problem. Using our decoup ling technique, we are further able to obtain fast convergence rates for strongl y con-vex pairwise loss functions. We are also able to analyze a class of memory efficient on-line learning algorithms for pairwise learning problems that use o nly a bounded subset of past training samples to update the hypothesis at each s tep. Finally, in order to complement our generalization bounds, we propose a nov el memory efficient online learning algorithm for higher order learning problems with bounded regret guarantees.
****
Non-Linear Stationary Subspace Analysis with Application to Video Classification

Mahsa Baktashmotlagh, Mehrtash Harandi, Abbas Bigdeli, Brian Lovell, Mathieu Salzmann

Low-dimensional representations are key to the success of many video classification algorithms. However, the commonly-used dimensionality reduction techniques fail to account for the fact that only part of the signal is shared across all the videos in one class. As a consequence, the resulting representations contain instance-specific information, which introduces noise in the classification process. In this paper, we introduce Non-Linear Stationary Subspace Analysis: A method that overcomes this issue by explicitly separating the stationary parts of the video signal (i.e., the parts shared across all videos in one class), from its non-stationary parts (i.e., specific to individual videos). We demonstrate the effectiveness of our approach on action recognition, dynamic texture classification and scene recognition.

****

Two-Sided Exponential Concentration Bounds for Bayes Error Rate and Shannon Entropy

Jean Honorio, Jaakkola Tommi

We provide a method that approximates the Bayes error rate and the Shannon entropy with high probability. The Bayes error rate approximation makes possible to build a classifier that polynomially approaches Bayes error rate. The Shannon entropy approximation provides provable performance guarantees for learning trees and Bayesian networks from continuous variables. Our results rely on some reasonable regularity conditions of the unknown probability distributions, and apply to bounded as well as unbounded variables.

****

That was fast! Speeding up NN search of high dimensional distributions.

Emanuele Coviello, Adeel Mumtaz, Antoni Chan, Gert Lanckriet

We present a data structure for fast nearest neighbor retrieval of generative models of documents based on KL divergence. Our data structure, which shares some similarity with Bregman Ball Trees, consists of a hierarchical partition of a database, and uses a novel branch and bound methodology for search. The main technical contribution of the paper is a novel and efficient algorithm for deciding whether to explore nodes during backtracking, based on a variational approximation. This reduces the number of computations per node, and overcomes the limitations of Bregman Ball Trees on high dimensional data. In addition, our strategy is applicable also to probability distributions with hidden state variables, and is not limited to regular exponential family distributions. Experiments demonstrate substantial speed-ups over both Bregman Ball Trees and over brute force search, on both moderate and high dimensional histogram data. In addition, experiments on linear dynamical systems demonstrate the flexibility of our approach to latent variable models.

****

Entropic Affinities: Properties and Efficient Numerical Computation

Max Vladymyrov, Miguel Carreira-Perpinan

Gaussian affinities are commonly used in graph-based methods such as spectral clustering or nonlinear embedding. Hinton and Roweis (2003) introduced a way to set the scale individually for each point so that it has a distribution over neighbors with a desired perplexity, or effective number of neighbors. This gives very good affinities that adapt locally to the data but are harder to compute. We study the mathematical properties of these “entropic affinities” and show that they implicitly define a continuously differentiable function in the input space and give bounds for it. We then devise a fast algorithm to compute the widths and affinities, based on robustified, quickly convergent root-finding methods combined with a tree- or density-based initialization scheme that exploits the slowly-varying behavior of this function. This algorithm is nearly optimal and much more accurate and fast than the existing bisection-based approach, particularly with large datasets, as we show with image and text data.

****

Local Deep Kernel Learning for Efficient Non-linear SVM Prediction

Cijo Jose, Prasoon Goyal, Parv Aggrwal, Manik Varma

Our objective is to speed up non-linear SVM prediction while maintaining classif ication accuracy above an acceptable limit. We generalize Localized Multiple Ker nel Learning so as to learn a primal feature space embedding which is high dimen sional, sparse and computationally deep. Primal based classification decouples p rediction costs from the number of support vectors and our tree-structured featu res efficiently encode non-linearities while speeding up prediction exponentiall y over the state-of-the-art. We develop routines for optimizing over the space o f tree-structured features and efficiently scale to problems with over half a mi llion training points. Experiments on benchmark data sets reveal that our formul ation can reduce prediction costs by more than three orders of magnitude in some cases with a moderate sacrifice in classification accuracy as compared to RBF-S VMs. Furthermore, our formulation leads to much better classification accuracies over leading methods.
****
Temporal Difference Methods for the Variance of the Reward To Go
Aviv Tamar, Dotan Di Castro, Shie Mannor
In this paper we extend temporal difference policy evaluation algorithms to perf ormance criteria that include the variance of the cumulative reward. Such criter ia are useful for risk management, and are important in domains such as finance and process control. We propose variants of both TD(0) and LSTD(λ) with linear f unction approximation, prove their convergence, and demonstrate their utility in a 4-dimensional continuous state space problem.
****
\proptoSVM for Learning with Label Proportions
Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, Shih-Fu Chang
We study the problem of learning with label proportions in which the training da ta is provided in groups and only the proportion of each class in each group is known. We propose a new method called proportion-SVM, or \proptoSVM, which expli citly models the latent unknown instance labels together with the known group la bel proportions in a large-margin framework. Unlike the existing works, our appr oach avoids making restrictive assumptions about the data. The \proptoSVM model leads to a non-convex integer programming problem. In order to solve it efficien tly, we propose two algorithms: one based on simple alternating optimization and the other based on a convex relaxation. Extensive experiments on standard datas ets show that \proptoSVM outperforms the state-of-the-art, especially for larger group sizes.
****
Parameter Learning and Convergent Inference for Dense Random Fields
Philipp Kraehenbuehl, Vladlen Koltun
Dense random fields are models in which all pairs of variables are directly conn ected by pairwise potentials. It has recently been shown that mean field inferen ce in dense random fields can be performed efficiently and that these models ena ble significant accuracy gains in computer vision applications. However, paramet er estimation for dense random fields is still poorly understood. In this paper, we present an efficient algorithm for learning parameters in dense random field s. All parameters are estimated jointly, thus capturing dependencies between the m. We show that gradients of a variety of loss functions over the mean field mar ginals can be computed efficiently. The resulting algorithm learns parameters th at directly optimize the performance of mean field inference in the model. As a supporting result, we present an efficient inference algorithm for dense random fields that is guaranteed to converge.
****
Loss-Proportional Subsampling for Subsequent ERM
Paul Mineiro, Nikos Karampatziakis
We propose a sampling scheme suitable for reducing a data set prior to selectin g a hypothesis with minimum empirical risk. The sampling only considers a subs et of the ultimate (unknown) hypothesis set, but can nonetheless guarantee that the final excess risk will compare favorably with utilizing the entire origina l data set. We demonstrate the practical benefits of our approach on a large da taset which we subsample and subsequently fit with boosted trees.

****
Scalable Simple Random Sampling and Stratified Sampling
Xiangrui Meng
Analyzing data sets of billions of records has now become a regular task in m
any companies and institutions. In the statistical analysis of those massive
data sets, sampling generally plays a very important role. In this work, w
e describe a scalable simple random sampling algorithm, named ScaSRS, which u
ses probabilistic thresholds to decide on the fly whether to accept, reject,
or wait-list an item independently of others. We prove, with high probability
, it succeeds and needs only O(\sqrtk) storage, where k is the sample size.
ScaSRS extends naturally to a scalable stratified sampling algorithm, which is
favorable for heterogeneous data sets. The proposed algorithms, when impl
emented in MapReduce, can effectively reduce the size of intermediate output
and greatly improve load balancing. Empirical evaluation on large-scale data
sets clearly demonstrates their superiority.
****
Riemannian Similarity Learning
Li Cheng
We consider a similarity-score based paradigm to address scenarios where either
the class labels are only partially revealed during learning, or the training an
d testing data are drawn from heterogeneous sources. The learning problem is sub
sequently formulated as optimization over a bilinear form of fixed rank. Our par
adigm bears similarity to metric learning, where the major difference lies in it
s aim of learning a rectangular similarity matrix, instead of a proper metric. W
e tackle this problem in a Riemannian optimization framework. In particular, we
consider its applications in pairwise-based action recognition, and cross-domain
image-based object recognition. In both applications, the proposed algorithm pr
oduces competitive performance on respective benchmark datasets.
****
On Compact Codes for Spatially Pooled Features
Yangqing Jia, Oriol Vinyals, Trevor Darrell
Feature encoding with an overcomplete dictionary has demonstrated good performan
ce in many applications, especially computer vision. In this paper we analyze th
e classification accuracy with respect to dictionary size by linking the encodin
g stage to kernel methods and \nystrom sampling, and obtain useful bounds on acc
uracy as a function of size. The \nystrom method also inspires us to revisit dic
tionary learning from local patches, and we propose to learn the dictionary in a
n end-to-end fashion taking into account pooling, a common computational layer i
n vision. We validate our contribution by showing how the derived bounds are abl
e to explain the observed behavior of multiple datasets, and show that the pooli
ng aware method efficiently reduces the dictionary size by a factor of two for a
given accuracy.
****
Dynamic Covariance Models for Multivariate Financial Time Series
Yue Wu, Jose Miguel Hernandez-Lobato, Ghahramani Zoubin
The accurate prediction of time-changing covariances is an important problem in
the modeling of multivariate financial data. However, some of the most popular m
odels suffer from a) overfitting problems and multiple local optima, b) failure
to capture shifts in market conditions and c) large computational costs. To addr
ess these problems we introduce a novel dynamic model for time-changing covarian
ces. Over-fitting and local optima are avoided by following a Bayesian approach
instead of computing point estimates. Changes in market conditions are captured
by assuming a diffusion process in parameter values, and finally computationally
efficient and scalable inference is performed using particle filters. Experimen
ts with financial data show excellent performance of the proposed method with re
spect to current standard models.
****
Revisiting the Nystrom method for improved large-scale machine learning
Alex Gittens, Michael Mahoney
We reconsider randomized algorithms for the low-rank approximation of SPSD matri

ces such as Laplacian and kernel matrices that arise in data analysis and machine learning applications. Our main results consist of an empirical evaluation of the performance quality and running time of sampling and projection methods on a diverse suite of SPSD matrices. Our results highlight complementary aspects of sampling versus projection methods, and they point to differences between uniform and nonuniform sampling methods based on leverage scores. We complement our empirical results with a suite of worst-case theoretical bounds for both random sampling and random projection methods. These bounds are qualitatively superior to existing bounds— e.g., improved additive-error bounds for spectral and Frobenius norm error and relative-error bounds for trace norm error.
****
Infinite Positive Semidefinite Tensor Factorization for Source Separation of Mixture Signals
Kazuyoshi Yoshii, Ryota Tomioka, Daichi Mochihashi, Masataka Goto
This paper presents a new class of tensor factorization called positive semidefinite tensor factorization (PSDTF) that decomposes a set of positive semidefinite (PSD) matrices into the convex combinations of fewer PSD basis matrices. PSDTF can be viewed as a natural extension of nonnegative matrix factorization. One of the main problems of PSDTF is that an appropriate number of bases should be given in advance. To solve this problem, we propose a nonparametric Bayesian model based on a gamma process that can instantiate only a limited number of necessary bases from the infinitely many bases assumed to exist. We derive a variational Bayesian algorithm for closed-form posterior inference and a multiplicative update rule for maximum-likelihood estimation. We evaluated PSDTF on both synthetic data and real music recordings to show its superiority.
****
A Unified Robust Regression Model for Lasso-like Algorithms
Wenzhuo Yang, Huan Xu
We develop a unified robust linear regression model and show that it is equivalent to a general regularization framework to encourage sparse-like structure that contains group Lasso and fused Lasso as specific examples. This provides a robustness interpretation of these widely applied Lasso-like algorithms, and allows us to construct novel generalizations of Lasso-like algorithms by considering different uncertainty sets. Using this robustness interpretation, we present new sparsity results, and establish the statistical consistency of the proposed regularized linear regression. This work extends a classical result from Xu et al. (2010) that relates standard Lasso with robust linear regression to learning problems with more general sparse-like structures, and provides new robustness-based tools to to understand learning problems with sparse-like structures.
****
Quickly Boosting Decision Trees – Pruning Underachieving Features Early
Ron Appel, Thomas Fuchs, Piotr Dollar, Pietro Perona
Boosted decision trees are one of the most popular and successful learning techniques used today. While exhibiting fast speeds at test time, relatively slow training makes them impractical for applications with real-time learning requirements. We propose a principled approach to overcome this drawback. We prove a bound on the error of a decision stump given its preliminary error on a subset of the training data; the bound may be used to prune unpromising features early on in the training process. We propose a fast training algorithm that exploits this bound, yielding speedups of an order of magnitude at no cost in the final performance of the classifier. Our method is not a new variant of Boosting; rather, it may be used in conjunction with existing Boosting algorithms and other sampling heuristics to achieve even greater speedups.
****
On the Statistical Consistency of Algorithms for Binary Classification under Class Imbalance
Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, Sanjay Chawla
Class imbalance situations, where one class is rare compared to the other, arise frequently in machine learning applications. It is well known that the usual misclassification error is ill-suited for measuring performance in such settings.

A wide range of performance measures have been proposed for this problem, in machine learning as well as in data mining, artificial intelligence, and various applied fields. However, despite the large number of studies on this problem, little is understood about the statistical consistency of the algorithms proposed with respect to the performance measures of interest. In this paper, we study consistency with respect to one such performance measure, namely the arithmetic mean of the true positive and true negative rates (AM), and establish that some simple methods that have been used in practice, such as applying an empirically determined threshold to a suitable class probability estimate or performing an empirically balanced form of risk minimization, are in fact consistent with respect to the AM (under mild conditions on the underlying distribution). Our results employ balanced losses that have been used recently in analyses of ranking problems (Kotlowski et al., 2011) and build on recent results on consistent surrogates for cost-sensitive losses (Scott, 2012). Experimental results confirm our consistency theorems.
****
Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment
Jason Chuang, Sonal Gupta, Christopher Manning, Jeffrey Heer
The use of topic models to analyze domain-specific texts often requires manual validation of the latent topics to ensure they are meaningful. We introduce a framework to support large-scale assessment of topical relevance. We measure the correspondence between a set of latent topics and a set of reference concepts to quantify four types of topical misalignment: junk, fused, missing, and repeated topics. Our analysis compares 10,000 topic model variants to 200 expert-provided domain concepts, and demonstrates how our framework can inform choices of model parameters, inference algorithms, and intrinsic measures of topical quality.
****
Online Kernel Learning with a Near Optimal Sparsity Bound
Lijun Zhang, Jinfeng Yi, Rong Jin, Ming Lin, Xiaofei He
In this work, we focus on Online Sparse Kernel Learning that aims to online learn a kernel classifier with a bounded number of support vectors. Although many online learning algorithms have been proposed to learn a sparse kernel classifier, most of them fail to bound the number of support vectors used by the final solution which is the average of the intermediate kernel classifiers generated by online algorithms. The key idea of the proposed algorithm is to measure the difficulty in correctly classifying a training example by the derivative of a smooth loss function, and give a more chance to a difficult example to be a support vector than an easy one via a sampling scheme. Our analysis shows that when the loss function is smooth, the proposed algorithm yields similar performance guarantee as the standard online learning algorithm but with a near optimal number of support vectors (up to a poly(lnT) factor). Our empirical study shows promising performance of the proposed algorithm compared to the state-of-the-art algorithms for online sparse kernel learning.
****
Spectral Learning of Hidden Markov Models from Dynamic and Static Data
Tzu-Kuo Huang, Jeff Schneider
We develop spectral learning algorithms for Hidden Markov Models that learn not only from time series, or dynamic data but also static data drawn independently from the HMM’s stationary distribution. This is motivated by the fact that static, orderless snapshots are usually easier to obtain than time series in quite a few dynamic modeling tasks. Building on existing spectral learning algorithms, our methods solve convex optimization problems minimizing squared loss on the dynamic data plus a regularization term on the static data. Experiments on synthetic and real human activities data demonstrate better prediction by the proposed method than existing spectral algorithms.
****
Analogy-preserving Semantic Embedding for Visual Object Categorization
Sung Ju Hwang, Kristen Grauman, Fei Sha
In multi-class categorization tasks, knowledge about the classes’ semantic relationships can provide valuable information beyond the class labels themselves. H

owever, existing techniques focus on preserving the semantic distances between c
lasses (e.g., according to a given object taxonomy for visual recognition), limi
ting the influence to pairwise structures. We propose to model \emphanalogies t
hat reflect the relationships between multiple pairs of classes simultaneously,
in the form “p is to q, as r is to s"". We translate semantic analogies into hi
gher-order geometric constraints called \emphanalogical parallelograms, and use
them in a novel convex regularizer for a discriminatively learned label embeddin
g.   Furthermore, we show how to discover analogies from attribute-based class d
escriptions, and how to prioritize those likely to reduce inter-class confusion.
  Evaluating our Analogy-preserving Semantic Embedding (ASE) on two visual recog
nition datasets, we demonstrate clear improvements over existing approaches, bot
h in terms of recognition accuracy and analogy completion.
****
Algebraic classifiers: a generic approach to fast cross-validation, online train
ing, and parallel training
Michael Izbicki
We use abstract algebra to derive new algorithms for fast cross-validation, onli
ne learning, and parallel learning. To use these algorithms on a classification
 model, we must show that the model has appropriate algebraic structure. It is
easy to give algebraic structure to some models, and we do this explicitly for B
ayesian classifiers and a novel variation of decision stumps called HomStumps.
But not all classifiers have an obvious structure, so we introduce the Free HomT
rainer. This can be used to give a “generic” algebraic structure to any classif
ier. We use the Free HomTrainer to give algebraic structure to bagging and boos
ting. In so doing, we derive novel online and parallel algorithms, and present
the first fast cross-validation schemes for these classifiers.
****
Factorial Multi-Task Learning : A Bayesian Nonparametric Approach
Sunil Gupta, Dinh Phung, Svetha Venkatesh
Multi-task learning is a paradigm shown to improve the performance of related ta
sks through their joint learning. However, for real-world data, it is usually di
fficult to assess the task relatedness and joint learning with unrelated tasks m
ay lead to serious performance degradations. To this end, we propose a framework
 that groups the tasks based on their relatedness in a low dimensional subspace
and allows a varying degree of relatedness among tasks by sharing the subspace b
ases across the groups. This provides the flexibility of no sharing when two set
s of tasks are unrelated and partial/total sharing when the tasks are related. I
mportantly, the number of task-groups and the subspace dimensionality are automa
tically inferred from the data. This feature keeps the model beyond a specific s
et of parameters. To realize our framework, we present a novel Bayesian nonparam
etric prior that extends the traditional hierarchical beta process prior using a
 Dirichlet process to permit potentially infinite number of child beta processes
. We apply our model for multi-task regression and classification applications.
Experimental results using several synthetic and real-world datasets show the su
periority of our model to other recent state-of-the-art multi-task learning meth
ods.
****
Modeling Information Propagation with Survival Theory
Manuel Gomez-Rodriguez, Jure Leskovec, Bernhard Schölkopf
Networks provide a ‘skeleton’ for the spread of contagions, like, information, i
deas, behaviors and diseases. Many times networks over which contagions diffuse
are unobserved and need to be inferred. Here we apply survival theory to develop
 general additive and multiplicative risk models under which the network inferen
ce problems can be solved efficiently by exploiting their convexity. Our additiv
e risk model generalizes several existing network inference models. We show all
these models are particular cases of our more general model. Our multiplicative
model allows for modeling scenarios in which a node can either increase or decre
ase the risk of activation of another node, in contrast with previous approaches
, which consider only positive risk increments. We evaluate the performance of o
ur network inference algorithms on large synthetic and real cascade datasets, an

d show that our models are able to predict the length and duration of cascades in real data.
****
Better Rates for Any Adversarial Deterministic MDP
Ofer Dekel, Elad Hazan
We consider regret minimization in adversarial deterministic Markov Decision Processes (ADMDPs) with bandit feedback. We devise a new algorithm that pushes the state-of-the-art forward in two ways: First, it attains a regret of O(T^2/3) with respect to the best fixed policy in hindsight, whereas the previous best regret bound was O(T^3/4). Second, the algorithm and its analysis are compatible with any feasible ADMDP graph topology, while all previous approaches required additional restrictions on the graph topology.
****
ABC Reinforcement Learning
Christos Dimitrakakis, Nikolaos Tziortziotis
We introduce a simple, general framework for likelihood-free Bayesian reinforcement learning, through Approximate Bayesian Computation (ABC). The advantage is that we only require a prior distribution on a class of simulators. This is useful when a probabilistic model of the underlying process is too complex to formulate, but where detailed simulation models are available. ABC-RL allows the use of any Bayesian reinforcement learning technique in this case. It can be seen as an extension of simulation methods to both planning and inference. We experimentally demonstrate the potential of this approach in a comparison with LSPI. Finally, we introduce a theorem showing that ABC is sound.
****
Sharp Generalization Error Bounds for Randomly-projected Classifiers
Robert Durrant, Ata Kaban
We derive sharp bounds on the generalization error of a generic linear classifier trained by   empirical risk minimization on randomly-projected data. We make no restrictive assumptions   (such as sparsity or separability) on the data: Instead we use the fact that, in a classification setting, the question of interest is really ‘what is the effect of random projection on the predicted class labels?’ and we therefore derive the exact probability of ‘label flipping’ under Gaussian random projection in order to quantify this effect precisely in our bounds.
****
On learning parametric-output HMMs
Aryeh Kontorovich, Boaz Nadler, Roi Weiss
We present a novel approach to learning an HMM whose outputs are distributed according to a parametric family. This is done by \em decoupling the learning task into two steps: first estimating the output parameters, and then estimating the hidden states transition probabilities. The first step is accomplished by fitting a mixture model to the output stationary distribution. Given the parameters of this mixture model, the second step is formulated as the solution of an easily solvable convex quadratic program. We provide an error analysis for the estimated transition probabilities and show they are robust to small perturbations in the estimates of the mixture parameters. Finally, we support our analysis with some encouraging empirical results.
****
LDA Topic Model with Soft Assignment of Descriptors to Words
Daphna Weinshall, Gal Levi, Dmitri Hanukaev
The LDA topic model is being used to model corpora of documents that can be represented by bags of words. Here we extend the LDA model to deal with documents that are represented more naturally by bags of continuous descriptors. Given a finite dictionary of words which are generative models of descriptors, our extended LDA model allows for the soft assignment of descriptors to (many) dictionary words. We derive variational inference and parameter estimation procedures for the extended model, which closely resemble those obtained for the original model, with two important differences: First, the histogram of word counts is replaced by a histogram of pseudo word counts, or sums of responsibilities over all descriptors. Second, parameter estimation now depends on the average covariance matrix

between these pseudo-counts, reflecting the fact that with soft assignment words are not independent.   We use this approach to address novelty detection, where we seek to identify video events with low posterior probability. Video events are described by a generative dynamic texture model, from which we naturally derive a dictionary of generative words. Using a benchmark dataset for novelty detection, we show a very significant improvement in the detection of novel events when using our extended LDA model with soft assignment to words as against hard assignment (the original model), achieving state of the art novelty detection results.
****

On autoencoder scoring

Hanna Kamyshanska, Roland Memisevic

Autoencoders are popular feature learning models because they are conceptually simple, easy to train and allow for efficient inference and training. Recent work has shown how certain autoencoders can assign an unnormalized “score” to data which measures how well the autoencoder can represent the data. Scores are commonly computed by using training criteria that relate the autoencoder to a probabilistic model, such as the Restricted Boltzmann Machine. In this paper we show how an autoencoder can assign meaningful scores to data independently of training procedure and without reference to any probabilistic model, by interpreting it as a dynamical system. We discuss how, and under which conditions, running the dynamical system can be viewed as performing gradient descent in an energy function, which in turn allows us to derive a score via integration. We also show how one can combine multiple, unnormalized scores into a generative classifier.
****

Infinite Markov-Switching Maximum Entropy Discrimination Machines

Sotirios Chatzis

In this paper, we present a method that combines the merits of Bayesian nonparametrics, specifically stick-breaking priors, and large-margin kernel machines in the context of sequential data classification. The proposed model postulates a set of (theoretically) infinite interdependent large-margin classifiers as model components, that robustly capture local nonlinearity of complex data. The postulated large-margin classifiers are connected in the context of a Markov-switching construction that allows for capturing complex temporal dynamics in the modeled datasets. Appropriate stick-breaking priors are imposed over the component switching mechanism of our model to allow for data-driven determination of the optimal number of component large-margin classifiers, under a standard nonparametric Bayesian inference scheme. Efficient model training is performed under the maximum entropy discrimination (MED) framework, which integrates the large-margin principle with Bayesian posterior inference. We evaluate our method using several real-world datasets, and compare it to state-of-the-art alternatives.
****

A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

We provide a first PAC-Bayesian analysis for domain adaptation (DA) which arises when the learning and test distributions differ. It relies on a novel distribution pseudodistance based on a disagreement averaging. Using this measure, we derive a PAC-Bayesian DA bound for the stochastic Gibbs classifier. This bound has the advantage of being directly optimizable for any hypothesis space. We specialize it to linear classifiers, and design a learning algorithm which shows interesting results on a synthetic problem and on a popular sentiment annotation task. This opens the door to tackling DA tasks by making use of all the PAC-Bayesian tools.
****

Sparse PCA through Low-rank Approximations

Dimitris Papailiopoulos, Alexandros Dimakis, Stavros Korokythakis

We introduce a novel algorithm that computes the k-sparse principal component of a positive semidefinite matrix A. Our algorithm is combinatorial and operates

by examining a discrete set of special vectors lying in a low-dimensional eigen-subspace of A. We obtain provable approximation guarantees that depend on the spectral profile of the matrix: the faster the eigenvalue decay, the better the quality of our approximation. For example, if the eigenvalues of A follow a power-law decay, we obtain a polynomial-time approximation algorithm for any desired accuracy. We implement our algorithm and test it on multiple artificial and real data sets. Due to a feature elimination step, it is possible to perform sparse PCA on data sets consisting of millions of entries in a few minutes. Our experimental evaluation shows that our scheme is nearly optimal while finding very sparse vectors. We compare to the prior state of the art and show that our scheme matches or outperforms previous algorithms in all tested data sets.
****

Computation-Risk Tradeoffs for Covariance-Thresholded Regression

Dinah Shender, John Lafferty

We present a family of linear regression estimators that provides a fine-grained tradeoff between statistical accuracy and computational efficiency. The estimators are based on hard thresholding of the sample covariance matrix entries together with l2-regularizion(ridge regression). We analyze the predictive risk of this family of estimators as a function of the threshold and regularization parameter. With appropriate parameter choices, the estimate is the solution to a sparse, diagonally dominant linear system, solvable in near-linear time. Our analysis shows how the risk varies with the sparsity and regularization level, thus establishing a statistical estimation setting for which there is an explicit, smooth tradeoff between risk and computation. Simulations are provided to support the theoretical analyses.
****

Exact Rule Learning via Boolean Compressed Sensing

Dmitry Malioutov, Kush Varshney

We propose an interpretable rule-based classification system based on ideas from Boolean compressed sensing. We represent the problem of learning individual conjunctive clauses or individual disjunctive clauses as a Boolean group testing problem, and apply a novel linear programming relaxation to find solutions. We derive results for exact rule recovery which parallel the conditions for exact recovery of sparse signals in the compressed sensing literature: although the general rule recovery problem is NP-hard, under some conditions on the Boolean ‘sensing’ matrix, the rule can be recovered exactly. This is an exciting development in rule learning where most prior work focused on heuristic solutions. Furthermore we construct rule sets from these learned clauses using set covering and boosting. We show competitive classification accuracy using the proposed approach.
****

Robust Sparse Regression under Adversarial Corruption

Yudong Chen, Constantine Caramanis, Shie Mannor

We consider high dimensional sparse regression with arbitrary – possibly, severe or coordinated – errors in the covariates matrix. We are interested in understanding how many corruptions we can tolerate, while identifying the correct support. To the best of our knowledge, neither standard outlier rejection techniques, nor recently developed robust regression algorithms (that focus only on corrupted response variables), nor recent algorithms for dealing with stochastic noise or erasures, can provide guarantees on support recovery. As we show, neither can the natural brute force algorithm that takes exponential time to find the subset of data and support columns, that yields the smallest regression error. We explore the power of a simple idea: replace the essential linear algebraic calculation – the inner product – with a robust counterpart that cannot be greatly affected by a controlled number of arbitrarily corrupted points: the trimmed inner product. We consider three popular algorithms in the uncorrupted setting: Thresholding Regression, Lasso, and the Dantzig selector, and show that the counterparts obtained using the trimmed inner product are provably robust.
****

Optimization with First-Order Surrogate Functions

Julien Mairal

In this paper, we study optimization methods consisting of iteratively minimizing surrogates of an objective function. By proposing several algorithmic variants and simple convergence analyses, we make two main contributions. First, we provide a unified viewpoint for several first-order optimization techniques such as accelerated proximal gradient, block coordinate descent, or Frank-Wolfe algorithms. Second, we introduce a new incremental scheme that experimentally matches or outperforms state-of-the-art solvers for large-scale optimization problems typically arising in machine learning.
****
Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation
Hema Koppula, Ashutosh Saxena
We consider the problem of detecting past activities as well as anticipating which activity will happen in the future and how. We start by modeling the rich spatio-temporal relations between human poses and objects (called affordances) using a conditional random field (CRF). However, because of the ambiguity in the temporal segmentation of the sub-activities that constitute an activity, in the past as well as in the future, multiple graph structures are possible. In this paper, we reason about these alternate possibilities by reasoning over multiple possible graph structures. We obtain them by approximating the graph with only additive features, which lends to efficient dynamic programming. Starting with this proposal graph structure, we then design moves to obtain several other likely graph structures. We then show that our approach improves the state-of-the-art significantly for detecting past activities as well as for anticipating future activities, on a dataset of 120 activity videos collected from four subjects.
****
Consistency versus Realizable H-Consistency for Multiclass Classification
Phil Long, Rocco Servedio
A consistent loss function for multiclass classification is one such that for any source of labeled examples, any tuple of scoring functions that minimizes the expected loss will have classification accuracy close to that of the Bayes optimal classifier. While consistency has been proposed as a desirable property for multiclass loss functions, we give experimental and theoretical results exhibiting a sequence of linearly separable data sources with the following property: a multiclass classification algorithm which optimizes a loss function due to Crammer and Singer (which is known not to be consistent) produces classifiers whose expected error goes to 0, while the expected error of an algorithm which optimizes a generalization of the loss function used by LogitBoost (a loss function which is known to be consistent) is bounded below by a positive constant.   We identify a property of a loss function, realizable consistency with respect to a restricted class of scoring functions, that accounts for this difference.   As our main technical results we show that the Crammer–Singer loss function is realizable consistent for the class of linear scoring functions, while the generalization of LogitBoost is not. Our result for LogitBoost is a special case of a more general theorem that applies to several other loss functions that have been proposed for multiclass classification.
****
Feature Multi-Selection among Subjective Features
Sivan Sabato, Adam Kalai
When dealing with subjective, noisy, or otherwise nebulous features, the “wisdom of crowds” suggests that one may benefit from multiple judgments of the same feature on the same object. We give theoretically-motivated ""feature multi-selection"" algorithms that choose, among a large set of candidate features, not only which features to judge but how many times to judge each one. We demonstrate the effectiveness of this approach for linear regression on a crowdsourced learning task of predicting people’s height and weight from photos, using features such as ""gender"" and ""estimated weight"" as well as culturally fraught ones such as ""attractive"".
****
Domain Adaptation under Target and Conditional Shift

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, Zhikun Wang
Let X denote the feature and Y the target. We consider domain adaptation under three possible scenarios: (1) the marginal P_Y changes, while the conditional P_X|Y stays the same (\it target shift), (2) the marginal P_Y is fixed, while the conditional P_X|Y changes with certain constraints (\it conditional shift), and (3) the marginal P_Y changes, and the conditional P_X|Y changes with constraints (\it generalized target shift). Using background knowledge, causal interpretations allow us to determine the correct situation for a problem at hand. We exploit importance reweighting or sample transformation to find the learning machine that works well on test data, and propose to estimate the weights or transformations by \it reweighting or transforming training data to reproduce the covariate distribution on the test domain. Thanks to kernel embedding of conditional as well as marginal distributions, the proposed approaches avoid distribution estimation, and are applicable for high-dimensional problems. Numerical evaluations on synthetic and real-world datasets demonstrate the effectiveness of the proposed framework.
****
Collective Stability in Structured Prediction: Generalization from One Example
Ben London, Bert Huang, Ben Taskar, Lise Getoor
Structured predictors enable joint inference over multiple interdependent output variables. These models are often trained on a small number of examples with large internal structure. Existing distribution-free generalization bounds do not guarantee generalization in this setting, though this contradicts a large body of empirical evidence from computer vision, natural language processing, social networks and other fields. In this paper, we identify a set of natural conditions – weak dependence, hypothesis complexity and a new measure, collective stability – that are sufficient for generalization from even a single example, without imposing an explicit generative model of the data. We then demonstrate that the complexity and stability conditions are satisfied by a broad class of models, including marginal inference in templated graphical models. We thus obtain uniform convergence rates that can decrease significantly faster than previous bounds, particularly when each structured example is sufficiently large and the number of training examples is constant, even one.
****
Stable Coactive Learning via Perturbation
Karthik Raman, Thorsten Joachims, Pannaga Shivaswamy, Tobias Schnabel
Coactive Learning is a model of interaction between a learning system (e.g. search engine) and its human users, wherein the system learns from (typically implicit) user feedback during operational use. User feedback takes the form of preferences, and recent work has introduced online algorithms that learn from this weak feedback. However, we show that these algorithms can be unstable and ineffective in real-world settings where biases and noise in the feedback are significant. In this paper, we propose the first coactive learning algorithm that can learn robustly despite bias and noise. In particular, we explore how presenting users with slightly perturbed objects (e.g., rankings) can stabilize the learning process. We theoretically validate the algorithm by proving bounds on the average regret. We also provide extensive empirical evidence on benchmarks and from a live search engine user study, showing that the new algorithm substantially outperforms existing methods.
****
Max-Margin Multiple-Instance Dictionary Learning
Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, Zhuowen Tu
Dictionary learning has became an increasingly important task in machine learning, as it is fundamental to the representation problem. A number of emerging techniques specifically include a codebook learning step, in which a critical knowledge abstraction process is carried out. Existing approaches in dictionary (codebook) learning are either generative (unsupervised e.g. k-means) or discriminative (supervised e.g. extremely randomized forests). In this paper, we propose a multiple instance learning (MIL) strategy (along the line of weakly supervised learning) for dictionary learning. Each code is represented by a classifier, such a

s a linear SVM, which naturally performs metric fusion for multi-channel features. We design a formulation to simultaneously learn mixtures of codes by maximizing classification margins in MIL. State-of-the-art results are observed in image classification benchmarks based on the learned codebooks, which observe both compactness and effectiveness.
****
Fast Semidifferential-based Submodular Function Optimization
Rishabh Iyer, Stefanie Jegelka, Jeff Bilmes
We present a practical and powerful new framework for both unconstrained and constrained submodular function optimization based on discrete semidifferentials (sub- and super-differentials). The resulting algorithms, which repeatedly compute and then efficiently optimize submodular semigradients, offer new and generalize many old methods for submodular optimization. Our approach, moreover, takes steps towards providing a unifying paradigm applicable to both submodular minimization and maximization, problems that historically have been treated quite distinctly. The practicality of our algorithms is important since interest in submodularity, owing to its natural and wide applicability, has recently been in ascendance within machine learning. We analyze theoretical properties of our algorithms for minimization and maximization, and show that many state-of-the-art maximization algorithms are special cases. Lastly, we complement our theoretical analyses with supporting empirical experiments.
****
Kernelized Bayesian Matrix Factorization
Mehmet Gönen, Suleiman Khan, Samuel Kaski
We extend kernelized matrix factorization with a fully Bayesian treatment and with an ability to work with multiple side information sources expressed as different kernels. Kernel functions have been introduced to matrix factorization to integrate side information about the rows and columns (e.g., objects and users in recommender systems), which is necessary for making out-of-matrix (i.e., cold start) predictions. We discuss specifically bipartite graph inference, where the output matrix is binary, but extensions to more general matrices are straightforward. We extend the state of the art in two key aspects: (i) A fully conjugate probabilistic formulation of the kernelized matrix factorization problem enables an efficient variational approximation, whereas fully Bayesian treatments are not computationally feasible in the earlier approaches. (ii) Multiple side information sources are included, treated as different kernels in multiple kernel learning that additionally reveals which side information sources are informative. Our method outperforms alternatives in predicting drug-protein interactions on two data sets. We then show that our framework can also be used for solving multilabel learning problems by considering samples and labels as the two domains where matrix factorization operates on. Our algorithm obtains the lowest Hamming loss values on 10 out of 14 multilabel classification data sets compared to five state-of-the-art multilabel learning algorithms.
****
Learning the Structure of Sum-Product Networks
Robert Gens, Domingos Pedro
Sum-product networks (SPNs) are a new class of deep probabilistic models. SPNs can have unbounded treewidth but inference in them is always tractable. An SPN is either a univariate distribution, a product of SPNs over disjoint variables, or a weighted sum of SPNs over the same variables. We propose the first algorithm for learning the structure of SPNs that takes full advantage of their expressiveness. At each step, the algorithm attempts to divide the current variables into approximately independent subsets. If successful, it returns the product of recursive calls on the subsets; otherwise it returns the sum of recursive calls on subsets of similar instances from the current training set. A comprehensive empirical study shows that the learned SPNs are typically comparable to graphical models in likelihood but superior in inference speed and accuracy.
****
Quantile Regression for Large-scale Applications
Jiyan Yang, Xiangrui Meng, Michael Mahoney

Quantile regression is a method to estimate the quantiles of the conditional distribution of a response variable, and as such it permits a much more accurate portrayal of the relationship between the response variable and observed covariates than methods such as Least-squares or Least Absolute Deviations regression. It can be expressed as a linear program, and interior-point methods can be used to find a solution for moderately large problems. Dealing with very large problems, \emphe.g., involving data up to and beyond the terabyte regime, remains a challenge. Here, we present a randomized algorithm that runs in time that is nearly linear in the size of the input and that, with constant probability, computes a (1+ε) approximate solution to an arbitrary quantile regression problem. Our algorithm computes a low-distortion subspace-preserving embedding with respect to the loss function of quantile regression. Our empirical evaluation illustrates that our algorithm is competitive with the best previous work on small to medium-sized problems, and that it can be implemented in MapReduce-like environments and applied to terabyte-sized problems.
****
Robust Regression on MapReduce
Xiangrui Meng, Michael Mahoney
Although the MapReduce framework is now the \emphde facto standard for analyzing massive data sets, many algorithms (in particular, many iterative algorithms popular in machine learning, optimization, and linear algebra) are hard to fit into MapReduce. Consider, \emphe.g., the \ell_p regression problem: given a matrix A ∈\mathbbR^m \times n and a vector b ∈\mathbbR^m, find a vector x^* ∈\mathbbR^n that minimizes f(x) = \|A x - b\|_p. The widely-used \ell_2 regression, \emphi.e., linear least-squares, is known to be highly sensitive to outliers; and choosing p ∈[1, 2) can help improve robustness. In this work, we propose an efficient algorithm for solving strongly over-determined (m ■n) robust \ell_p regression problems to moderate precision on MapReduce. Our empirical results on data up to the terabyte scale demonstrate that our algorithm is a significant improvement over traditional iterative algorithms on MapReduce for \ell_1 regression, even for a fairly small number of iterations. In addition, our proposed interior-point cutting-plane method can also be extended to solving more general convex problems on MapReduce.
****
Infinitesimal Annealing for Training Semi-Supervised Support Vector Machines
Kohei Ogawa, Motoki Imamura, Ichiro Takeuchi, Masashi Sugiyama
The semi-supervised support vector machine (S3VM) is a maximum-margin classification algorithm based on both labeled and unlabeled data. Training S3VM involves either a combinatorial or non-convex optimization problem and thus finding the global optimal solution is intractable in practice. It has been demonstrated that a key to successfully find a good (local) solution of S3VM is to gradually increase the effect of unlabeled data, a la annealing. However, existing algorithms suffer from the trade-off between the resolution of annealing steps and the computation cost. In this paper, we go beyond this trade-off by proposing a novel training algorithm that efficiently performs annealing with an infinitesimal resolution. Through experiments, we demonstrate that the proposed infinitesimal annealing algorithm tends to produce better solutions with less computation time than existing approaches.
****
One-Pass AUC Optimization
Wei Gao, Rong Jin, Shenghuo Zhu, Zhi-Hua Zhou
AUC is an important performance measure and many algorithms have been devoted to AUC optimization, mostly by minimizing a surrogate convex loss on a training data set. In this work, we focus on one-pass AUC optimization that requires only going through the training data once without storing the entire training dataset, where conventional online learning algorithms cannot be applied directly because AUC is measured by a sum of losses defined over pairs of instances from different classes. We develop a regression-based algorithm which only needs to maintain the first and second order statistics of training data in memory, resulting a storage requirement independent from the size of training data. To efficiently h

andle high dimensional data, we develop a randomized algorithm that approximates
 the covariance matrices by low rank matrices. We verify, both theoretically and
 empirically, the effectiveness of the proposed algorithm.
****
Learning Convex QP Relaxations for Structured Prediction
Jeremy Jancsary, Sebastian Nowozin, Carsten Rother
We introduce a new large margin approach to discriminative training of intractab
le discrete graphical models. Our approach builds on a convex quadratic programm
ing relaxation of the MAP inference problem. The model parameters are trained di
rectly within this restricted class of energy functions so as to optimize the pr
edictions on the training data. We address the issue of how to parameterize the
resulting model and point out its relation to existing approaches. The primary m
otivation behind our use of the QP relaxation is its computational efficiency; y
et, empirically, its predictive accuracy compares favorably to more expensive ap
proaches. This makes it an appealing choice for many practical tasks.
****
Concurrent Reinforcement Learning from Customer Interactions
David Silver, Leonard Newnham, David Barker, Suzanne Weller, Jason McFall
In this paper, we explore applications in which a company interacts concurrently
 with many customers. The company has an objective function, such as maximising
revenue, customer satisfaction, or customer loyalty, which depends primarily on
the sequence of interactions between company and customer. A key aspect of this
setting is that interactions with different customers occur in parallel. As a re
sult, it is imperative to learn online from partial interaction sequences, so th
at information acquired from one customer is efficiently assimilated and applied
 in subsequent interactions with other customers. We present the first framework
 for concurrent reinforcement learning, using a variant of temporal-difference l
earning to learn efficiently from partial interaction sequences.   We evaluate o
ur algorithms in two large-scale test-beds for online and email interaction resp
ectively, generated from a database of 300,000 customer records.
****
Saving Evaluation Time for the Decision Function in Boosting: Representation and
 Reordering Base Learner
Peng Sun, Jie Zhou
For a well trained Boosting classifier, we are interested in how to save the tes
ting time, i.e., to make the decision without evaluating all the base learners.
To address this problem, in previous work the base learners are sequentially cal
culated and early stopping is allowed if the decision function has been confiden
t enough to output its value. In such a chain structure, the order of base learn
ers is critical: better order can lead to less evaluation time.   In this paper
, we present a novel method for ordering. We base our discussion on the data str
ucture representing Boosting’s decision function. Viewing the decision function
a boolean expression, we propose a Binary Valued Tree for its representation. As
 a secondary contribution, such a representation unifies the work by previous re
searchers and helps devise new representation. Also, its connection to Binary De
cision Diagram(BDD) is discussed.
****
Stability and Hypothesis Transfer Learning
Ilja Kuzborskij, Francesco Orabona
We consider the transfer learning scenario, where the learner does not have acce
ss to the source domain directly, but rather operates on the basis of hypotheses
 induced from it – the Hypothesis Transfer Learning (HTL) problem. Particularly,
 we conduct a theoretical analysis of HTL by considering the algorithmic stabili
ty of a class of HTL algorithms based on Regularized Least Squares with biased r
egularization. We show that the relatedness of source and target domains acceler
ates the convergence of the Leave-One-Out error to the generalization error, thu
s enabling the use of the Leave-One-Out error to find the optimal transfer param
eters, even in the presence of a small training set. In case of unrelated domain
s we also suggest a theoretically principled way to prevent negative transfer, s
o that in the limit we recover the performance of the algorithm not using any kn

owledge from the source domain.
****
Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models
Mohammad Emtiyaz Khan, Aleksandr Aravkin, Michael Friedlander, Matthias Seeger
Latent Gaussian models (LGMs) are widely used in statistics and machine learning. Bayesian inference in non-conjugate LGM is difficult due to intractable integrals involving the Gaussian prior and non-conjugate likelihoods. Algorithms based on Variational Gaussian (VG) approximations are widely employed since they strike a favorable balance between accuracy, generality, speed, and ease of use. However, the structure of optimization problems associated with them remains poorly understood, and standard solvers take too long to converge. In this paper, we derive a novel dual variational inference approach, which exploits the convexity property of the VG approximations. The implications of our approach is that we obtain an algorithm that solves a convex optimization problem, reduces the number of variational parameters, and converges much faster than previous methods. Using real world data, we demonstrate these advantages on a variety of LGMs including Gaussian process classification and latent Gaussian Markov random fields.
****
Modeling Temporal Evolution and Multiscale Structure in Networks
Tue Herlau, Morten Mørup, Mikkel Schmidt
Many real-world networks exhibit both temporal evolution and multiscale structure. We propose a model for temporally correlated multifurcating hierarchies in complex networks which jointly capture both effects. We use the Gibbs fragmentation tree as prior over multifurcating trees and a change-point model to account for the temporal evolution of each vertex. We demonstrate that our model is able to infer time-varying multiscale structure in synthetic as well as three real world time-evolving complex networks. Our modeling of the temporal evolution of hierarchies brings new insights into the changing roles and position of entities and possibilities for better understanding these dynamic complex systems.
****
Dependent Normalized Random Measures
Changyou Chen, Vinayak Rao, Wray Buntine, Yee Whye Teh
In this paper we propose two constructions of dependent normalized random measures, a class of nonparametric priors over dependent probability measures. Our constructions, which we call mixed normalized random measures (MNRM) and thinned normalized random measures (TNRM), involve (respectively) weighting and thinning parts of a shared underlying Poisson process before combining them together. We show that both MNRM and TNRM are marginally normalized random measures, resulting in well understood theoretical properties. We develop marginal and slice samplers for both models, the latter necessary for inference in TNRM. In time-varying topic modelling experiments, both models exhibit superior performance over related dependent models such as the hierarchical Dirichlet process and the spatial normalized Gamma process.
****
Fast Max-Margin Matrix Factorization with Data Augmentation
Minjie Xu, Jun Zhu, Bo Zhang
Existing max-margin matrix factorization (M3F) methods either are computationally inefficient or need a model selection procedure to determine the number of latent factors. In this paper we present a probabilistic M3F model that admits a highly efficient Gibbs sampling algorithm through data augmentation. We further extend our approach to incorporate Bayesian nonparametrics and build accordingly a truncation-free nonparametric M3F model where the number of latent factors is literally unbounded and inferred from data. Empirical studies on two large real-world data sets verify the efficacy of our proposed methods.
****
Natural Image Bases to Represent Neuroimaging Data
Ashish Gupta, Murat Ayhan, Anthony Maida
Visual inspection of neuroimagery is susceptible to human eye limitations. Computerized methods have been shown to be equally or more effective than human cl

inicians in diagnosing dementia from neuroimages. Nevertheless, much of the work involves the use of domain expertise to extract hand-crafted features. The key technique in this paper is the use of cross-domain features to represent MRI data. We used a sparse autoencoder to learn a set of bases from natural images and then applied convolution to extract features from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset.Using this new representation, we classify MRI instances into three categories: Alzheimer’s Disease (AD), Mild Cognitive Impairment (MCI) and Healthy Control (HC).Our approach, in spite of being very simple, achieved high classification performance, which is competitive with or better than other approaches.
****
Breaking the Small Cluster Barrier of Graph Clustering
Nir Ailon, Yudong Chen, Huan Xu
This paper investigates graph clustering in the planted cluster model in the presence of \em small clusters. Traditional results dictate that for an algorithm to provably correctly recover the clusters, \em all clusters must be sufficiently large (in particular, \tildeΩ(\sqrtn) where n is the number of nodes of the graph). We show that this is not really a restriction: by a more refined analysis of the trace-norm based matrix recovery approach proposed in (Jalali et al. 2011) and (Chen et al. 2012), we prove that small clusters, under certain mild assuptions, do not hinder recovery of large ones. Based on this result, we further devise an iterative algorithm to recover \em almost all clusters via a “peeling strategy”, i.e., recover large clusters first, leading to a reduced problem, and repeat this procedure. These results are extended to the \em partial observation setting, in which only a (chosen) part of the graph is observed. The peeling strategy gives rise to an active learning algorithm, in which edges adjacent to smaller clusters are queried more often as large clusters are learned (and removed). Our findings are supported by experiments. From a high level, this paper sheds novel insights on high-dimesional statistics and learning structured data, by presenting a structured matrix learning problem for which a one shot convex relaxation approach necessarily fails, but a carefully constructed sequence of convex relaxations does the job.
****
Approximate Inference in Collective Graphical Models
Daniel Sheldon, Tao Sun, Akshat Kumar, Tom Dietterich
We study the problem of approximate inference in collective graphical models (CGMs), which were recently introduced to model the problem of learning and inference with noisy aggregate observations. We first analyze the complexity of inference in CGMs: unlike inference in conventional graphical models, exact inference in CGMs is NP-hard even for tree-structured models. We then develop a tractable convex approximation to the NP-hard MAP inference problem in CGMs, and show how to use MAP inference for approximate marginal inference within the EM framework. We demonstrate empirically that these approximation techniques can reduce the computational cost of inference by two orders of magnitude and the cost of learning by at least an order of magnitude while providing solutions of equal or better quality.
****
Scaling the Indian Buffet Process via Submodular Maximization
Colorado Reed, Ghahramani Zoubin
Inference for latent feature models is inherently difficult as the inference space grows exponentially with the size of the input data and number of latent features. In this work, we use Kurihara & Wellings (2008)’s maximization-expectation framework to perform approximate MAP inference for linear-Gaussian latent feature models with an Indian Buffet Process (IBP) prior. This formulation yields a submodular function of the features that corresponds to a lower bound on the model evidence. By adding a constant to this function, we obtain a nonnegative submodular function that can be maximized via a greedy algorithm that obtains at least a 1/3-approximation to the optimal solution. Our inference method scales linearly with the size of the input data, and we show the efficacy of our method on the largest datasets currently analyzed using an IBP model.

****
Mini-Batch Primal and Dual Methods for SVMs
Martin Takac, Avleen Bijral, Peter Richtarik, Nati Srebro
We address the issue of using mini-batches in stochastic optimization of SVMs.
We show that the same quantity, the spectral norm of the data, controls the par
allelization speedup obtained for both primal stochastic subgradient descent(SG
D) and stochastic dual coordinate ascent (SCDA) methods and use it to derive nov
el variants of mini-batched SDCA. Our guarantees for both methods are expressed
in terms of the original nonsmooth primal problem based on the hinge-loss.
****
The lasso, persistence, and cross-validation
Darren Homrighausen, Daniel McDonald
During the last fifteen years, the lasso procedure has been the target of a subs
tantial amount of theoretical and applied research. Correspondingly, many result
s are known about its behavior for a fixed or optimally chosen smoothing paramet
er (given up to unknown constants). Much less, however, is known about the lasso
’s behavior when the smoothing parameter is chosen in a data dependent way. To t
his end, we give the first result about the risk consistency of lasso when the s
moothing parameter is chosen via cross-validation. We consider the high-dimensio
nal setting wherein the number of predictors p=n^α, α>0 grows with the number of
 observations.
****
Spectral Experts for Estimating Mixtures of Linear Regressions
Arun Tejasvi Chaganty, Percy Liang
Discriminative latent-variable models are typically learned using EM or gradient
-based optimization, which suffer from local optima. In this paper, we develop
a new computationally efficient and provably consistent estimator for the mixtur
e of linear regressions, a simple instance of discriminative latent-variable mod
els. Our approach relies on a low-rank linear regression to recover a symmetric
 tensor, which can be factorized into the parameters using the tensor power meth
od. We prove rates of convergence for our estimator and provide an empirical ev
aluation illustrating its strengths relative to local optimization (EM).
****
Distribution to Distribution Regression
Junier Oliva, Barnabas Poczos, Jeff Schneider
We analyze ’Distribution to Distribution regression’ where one is regressing a m
apping where both the covariate (inputs) and response (outputs) are distribution
s. No parameters on the input or output distributions are assumed, nor are any s
trong assumptions made on the measure from which input distributions are drawn f
rom. We develop an estimator and derive an upper bound for the L2 risk; also, we
 show that when the effective dimension is small enough (as measured by the doub
ling dimension), then the risk converges to zero with a polynomial rate.
****
Regularization of Neural Networks using DropConnect
Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, Rob Fergus
We introduce DropConnect, a generalization of DropOut, for regularizing large fu
lly-connected layers within neural networks. When training with Dropout, a rando
mly selected subset of activations are set to zero within each layer. DropConnec
t instead sets a randomly selected subset of weights within the network to zero.
 Each unit thus receives input from a random subset of units in the previous lay
er. We derive a bound on the generalization performance of both Dropout and Drop
Connect. We then evaluate DropConnect on a range of datasets, comparing to Dropo
ut, and show state-of-the-art results on several image recoginition benchmarks c
an be obtained by aggregating multiple DropConnect-trained models.
****
Gaussian Process Kernels for Pattern Discovery and Extrapolation
Andrew Wilson, Ryan Adams
Gaussian processes are rich distributions over functions, which provide a Bayesi
an nonparametric approach to smoothing and interpolation. We introduce simple c
losed form kernels that can be used with Gaussian processes to discover patterns

and enable extrapolation. These kernels are derived by modelling a spectral de nsity – the Fourier transform of a kernel – with a Gaussian mixture. The propos ed kernels support a broad class of stationary covariances, but Gaussian process inference remains simple and analytic. We demonstrate the proposed kernels by discovering patterns and performing long range extrapolation on synthetic exampl es, as well as atmospheric CO2 trends and airline passenger data. We also show that it is possible to reconstruct several popular standard covariances within o ur framework.
****
Anytime Representation Learning
Zhixiang Xu, Matt Kusner, Gao Huang, Kilian Weinberger
Evaluation cost during test-time is becoming increasingly important as many real -world applications need fast evaluation (e.g. web search engines, email spam fi ltering) or use expensive features (e.g. medical diagnosis). We introduce Anytim e Feature Representations (AFR), a novel algorithm that explicitly addresses thi s trade-off in the data representation rather than in the classifier. This enabl es us to turn conventional classifiers, in particular Support Vector Machines, i nto test-time cost sensitive anytime classifiers - combining the advantages of a nytime learning and large-margin classification.
****
Algorithms for Direct 0–1 Loss Optimization in Binary Classification
Tan Nguyen, Scott Sanner
While convex losses for binary classification are attractive due to the existenc e of numerous (provably) efficient methods for finding their global optima, they are sensitive to outliers. On the other hand, while the non-convex 0–1 loss is robust to outliers, it is NP-hard to optimize and thus rarely directly optimize d in practice. In this paper, however, we do just that: we explore a variety of practical methods for direct (approximate) optimization of the 0–1 loss based o n branch and bound search, combinatorial search, and coordinate descent on smoot h, differentiable relaxations of 0–1 loss. Empirically, we compare our proposed algorithms to logistic regression, SVM, and the Bayes point machine showing that the proposed 0–1 loss optimization algorithms perform at least as well and offe r a clear advantage in the presence of outliers. To this end, we believe this w ork reiterates the importance of 0–1 loss and its robustness properties while ch allenging the notion that it is difficult to directly optimize.
****
Top-k Selection based on Adaptive Sampling of Noisy Preferences
Robert Busa-Fekete, Balazs Szorenyi, Weiwei Cheng, Paul Weng, Eyke Huellermeier
We consider the problem of reliably selecting an optimal subset of fixed size fr om a given set of choice alternatives, based on noisy information about the qual ity of these alternatives. Problems of similar kind have been tackled by means o f adaptive sampling schemes called racing algorithms. However, in contrast to ex isting approaches, we do not assume that each alternative is characterized by a real-valued random variable, and that samples are taken from the corresponding d istributions. Instead, we only assume that alternatives can be compared in terms of pairwise preferences. We propose and formally analyze a general preference-b ased racing algorithm that we instantiate with three specific ranking procedures and corresponding sampling schemes. Experiments with real and synthetic data ar e presented to show the efficiency of our approach.
****
The Extended Parameter Filter
Yusuf Bugra Erol, Lei Li, Bharath Ramsundar, Russell Stuart
The parameters of temporal models, such as dynamic Bayesian networks, may be mod elled in a Bayesian context as static or atemporal variables that influence tran sition probabilities at every time step. Particle filters fail for models that i nclude such variables, while methods that use Gibbs sampling of parameter variab les may incur a per-sample cost that grows linearly with the length of the obser vation sequence. Storvik devised a method for incremental computation of exact s ufficient statistics that, for some cases, reduces the per-sample cost to a cons tant. In this paper, we demonstrate a connection between Storvik’s filter and a

Kalman filter in parameter space and establish more general conditions under wh ich Storvik’s filter works. Drawing on an analogy to the extended Kalman filter, we develop and analyze, both theoretically and experimentally, a Taylor approxi mation to the parameter posterior that allows Storvik’s method to be applied to a broader class of models. Our experiments on both synthetic examples and real a pplications show improvement over existing methods.
****
Exploiting Ontology Structures and Unlabeled Data for Learning
Nina Balcan, Avrim Blum, Yishay Mansour
We present and analyze a theoretical model designed to understand and explain t he effectiveness of ontologies for learning multiple related tasks from primari ly unlabeled data. We present both information-theoretic results as well as ef ficient algorithms. We show in this model that an ontology, which specifies t he relationships between multiple outputs, in some cases is sufficient to comp letely learn a classification using a large unlabeled data source.
****
O(logT) Projections for Stochastic Optimization of Smooth and Strongly Convex Fu nctions
Lijun Zhang, Tianbao Yang, Rong Jin, Xiaofei He
Traditional algorithms for stochastic optimization require projecting the soluti on at each iteration into a given domain to ensure its feasibility. When facing complex domains, such as the positive semidefinite cone, the projection operatio n can be expensive, leading to a high computational cost per iteration. In this paper, we present a novel algorithm that aims to reduce the number of projection s for stochastic optimization. The proposed algorithm combines the strength of s everal recent developments in stochastic optimization, including mini-batches, e xtra-gradient, and epoch gradient descent, in order to effectively explore the s moothness and strong convexity. We show, both in expectation and with a high pro bability, that when the objective function is both smooth and strongly convex, t he proposed algorithm achieves the optimal O(1/T) rate of convergence with only O(logT) projections. Our empirical study verifies the theoretical result.
****
Optimizing the F-Measure in Multi-Label Classification: Plug-in Rule Approach ve rsus Structured Loss Minimization
Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, E yke Huellermeier
We compare the plug-in rule approach for optimizing the F-measure in multi-label classification with an approach based on structured loss minimization, such as the structured support vector machine (SSVM). Whereas the former derives an opti mal prediction from a probabilistic model in a separate inference step, the latt er seeks to optimize the F-measure directly during the training phase. We introd uce a novel plug-in rule algorithm that estimates all parameters required for a Bayes-optimal prediction via a set of multinomial regression models, and we comp are this algorithm with SSVMs in terms of computational complexity and statistic al consistency. As a main theoretical result, we show that our plug-in rule algo rithm is consistent, whereas the SSVM approaches are not. Finally, we present re sults of a large experimental study showing the benefits of the introduced algor ithm.
****
On the importance of initialization and momentum in deep learning
Ilya Sutskever, James Martens, George Dahl, Geoffrey Hinton
Deep and recurrent neural networks (DNNs and RNNs respectively) are powerful mod els that were considered to be almost impossible to train using stochastic gradi ent descent with momentum. In this paper, we show that when stochastic gradient descent with momentum uses a well-designed random initialization and a particula r type of slowly increasing schedule for the momentum parameter, it can train bo th DNNs and RNNs (on datasets with long-term dependencies) to levels of performa nce that were previously achievable only with Hessian-Free optimization. We find that both the initialization and the momentum are crucial since poorly initiali zed networks cannot be trained with momentum and well-initialized networks perfo

rm markedly worse when the momentum is absent or poorly tuned. Our success training these models suggests that previous attempts to train deep and recurrent neural networks from random initializations have likely failed due to poor initialization schemes. Furthermore, carefully tuned momentum methods suffice for dealing with the curvature issues in deep and recurrent network training objectives without the need for sophisticated second-order methods.
****

A non-IID Framework for Collaborative Filtering with Restricted Boltzmann Machines

Kostadin Georgiev, Preslav Nakov

We propose a framework for collaborative filtering based on Restricted Boltzmann Machines (RBM), which extends previous RBM-based approaches in several important directions. First, while previous RBM research has focused on modeling the correlation between item ratings, we model both user-user and item-item correlations in a unified hybrid non-IID framework. We further use real values in the visible layer as opposed to multinomial variables, thus taking advantage of the natural order between user-item ratings. Finally, we explore the potential of combining the original training data with data generated by the RBM-based model itself in a bootstrapping fashion. The evaluation on two MovieLens datasets (with 100K and 1M user-item ratings, respectively), shows that our RBM model rivals the best previously-proposed approaches.
****

Intersecting singularities for multi-structured estimation

Emile Richard, Francis BACH, Jean-Philippe Vert

We address the problem of designing a convex nonsmooth regularizer encouraging multiple structural effects simultaneously. Focusing on the inference of sparse and low-rank matrices we suggest a new complexity index and a convex penalty approximating it. The new penalty term can be written as the trace norm of a linear function of the matrix. By analyzing theoretical properties of this family of regularizers we come up with oracle inequalities and compressed sensing results ensuring the quality of our regularized estimator. We also provide algorithms and supporting numerical experiments.
****

Structure Discovery in Nonparametric Regression through Compositional Kernel Search

David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, Ghahramani Zoubin

Despite its importance, choosing the structural form of the kernel in nonparametric regression remains a black art. We define a space of kernel structures which are built compositionally by adding and multiplying a small number of base kernels. We present a method for searching over this space of structures which mirrors the scientific discovery process. The learned structures can often decompose functions into interpretable components and enable long-range extrapolation on time-series datasets. Our structure search method outperforms many widely used kernels and kernel combination methods on a variety of prediction tasks.
****

Copy or Coincidence? A Model for Detecting Social Influence and Duplication Events

Lisa Friedland, David Jensen, Michael Lavine

In this paper, we analyze the task of inferring rare links between pairs of entities that seem too similar to have occurred by chance. Variations of this task appear in such diverse areas as social network analysis, security, fraud detection, and entity resolution. To address the task in a general form, we propose a simple, flexible mixture model in which most entities are generated independently from a distribution but a small number of pairs are constrained to be similar. We predict the true pairs using a likelihood ratio that trades off the entities' similarity with their rarity. This method always outperforms using only similarity; however, with certain parameter settings, similarity turns out to be surprisingly competitive. Using real data, we apply the model to detect twins given their birth weights and to re-identify cell phone users based on distinctive usage patterns.

****
Smooth Operators
Steffen Grunewalder, Gretton Arthur, John Shawe-Taylor
We develop a generic approach to form smooth versions of basic mathematical operations like multiplication, composition, change of measure, and conditional expectation, among others. Operations which result in functions outside the reproducing kernel Hilbert space (such as the product of two RKHS functions) are approximated via a natural cost function, such that the solution is guaranteed to be in the targeted RKHS. This approximation problem is reduced to a regression problem using an adjoint trick, and solved in a vector-valued RKHS, consisting of continuous, linear, smooth operators which map from an input, real-valued RKHS to the desired target RKHS. Important constraints, such as an almost everywhere positive density, can be enforced or approximated naturally in this framework, using convex constraints on the operators. Finally, smooth operators can be composed to accomplish more complex machine learning tasks, such as the sum rule and kernelized approximate Bayesian inference, where state-of-the-art convergence rates are obtained.
****
The Cross-Entropy Method Optimizes for Quantiles
Sergiu Goschin, Ari Weinstein, Michael Littman
Cross-entropy optimization (CE) has proven to be a powerful tool for search in control environments. In the basic scheme, a distribution over proposed solutions is repeatedly adapted by evaluating a sample of solutions and refocusing the distribution on a percentage of those with the highest scores. We show that, in the kind of noisy evaluation environments that are common in decision-making domains, this percentage-based refocusing does not optimize the expected utility of solutions, but instead a quantile metric. We provide a variant of CE (Proportional CE) that effectively optimizes the expected value. We show using variants of established noisy environments that Proportional CE can be used in place of CE and can improve solution quality.
****
Topic Discovery through Data Dependent and Random Projections
Weicong Ding, Mohammad Hossein Rohban, Prakash Ishwar, Venkatesh Saligrama
We present algorithms for topic modeling based on the geometry of cross-document word-frequency patterns. This perspective gains significance under the so called separability condition. This is a condition on existence of novel-words that are unique to each topic. We present a suite of highly efficient algorithms with provable guarantees based on data-dependent and random projections to identify novel words and associated topics. Our key insight here is that the maximum and minimum values of cross-document frequency patterns projected along any direction are associated with novel words. While our sample complexity bounds for topic recovery are similar to the state-of-art, the computational complexity of our random projection scheme scales linearly with the number of documents and the number of words per document. We present several experiments on synthetic and realworld datasets to demonstrate qualitative and quantitative merits of our scheme.
****
Bayesian Learning of Recursively Factored Environments
Marc Bellemare, Joel Veness, Michael Bowling
Model-based reinforcement learning techniques have historically encountered a number of difficulties scaling up to large observation spaces. One promising approach has been to decompose the model learning task into a number of smaller, more manageable sub-problems by factoring the observation space. Typically, many different factorizations are possible, which can make it difficult to select an appropriate factorization without extensive testing. In this paper we introduce the class of recursively decomposable factorizations, and show how exact Bayesian inference can be used to efficiently guarantee predictive performance close to the best factorization in this class. We demonstrate the strength of this approach by presenting a collection of empirical results for 20 different Atari 2600 games.
****

Selective sampling algorithms for cost-sensitive multiclass prediction
Alekh Agarwal
In this paper, we study the problem of active learning for cost-sensitive multic
lass classification. We propose selective sampling algorithms, which process th
e data in a streaming fashion, querying only a subset of the labels. For these
algorithms, we analyze the regret and label complexity when the labels are gener
ated according to a generalized linear model. We establish that the gains of act
ive learning over passive learning can range from none to exponentially large, b
ased on a natural notion of margin. We also present a safety guarantee to guard
 against model mismatch. Numerical simulations show that our algorithms indeed
obtain a low regret with a small number of queries.
****
The Bigraphical Lasso
Alfredo Kalaitzis, John Lafferty, Neil D. Lawrence, Shuheng Zhou
The i.i.d. assumption in machine learning is endemic, but often flawed. Complex
data sets exhibit partial correlations between both instances and features. A mo
del specifying both types of correlation can have a number of parameters that sc
ales quadratically with the number of features and data points. We introduce the
 bigraphical lasso, an estimator for precision matrices of matrix-normals based
on the Cartesian product of graphs. A prominent product in spectral graph theory
, this structure has appealing properties for regression, enhanced sparsity and
interpretability. To deal with the parameter explosion we introduce L1 penalties
 and fit the model through a flip-flop algorithm that results in a linear number
 of lasso regressions.
****
Almost Optimal Exploration in Multi-Armed Bandits
Zohar Karnin, Tomer Koren, Oren Somekh
We study the problem of exploration in stochastic Multi-Armed Bandits. Even in t
he simplest setting of identifying the best arm, there remains a logarithmic mul
tiplicative gap between the known lower and upper bounds for the number of arm p
ulls required for the task. This extra logarithmic factor is quite meaningful in
 nowadays large-scale applications. We present two novel, parameter-free algorit
hms for identifying the best arm, in two different settings: given a target conf
idence and given a target budget of arm pulls, for which we prove upper bounds w
hose gap from the lower bound is only doubly-logarithmic in the problem paramete
rs. We corroborate our theoretical results with experiments demonstrating that o
ur algorithm outperforms the state-of-the-art and scales better as the size of t
he problem increases.
****
Deep Canonical Correlation Analysis
Galen Andrew, Raman Arora, Jeff Bilmes, Karen Livescu
We introduce Deep Canonical Correlation Analysis (DCCA), a method to learn compl
ex nonlinear transformations of two views of data such that the resulting repres
entations are highly linearly correlated. Parameters of both transformations are
 jointly learned to maximize the (regularized) total correlation.   It can be vi
ewed as a nonlinear extension of the linear method \emphcanonical correlation an
alysis (CCA). It is an alternative to the nonparametric method \emphkernel cano
nical correlation analysis (KCCA) for learning correlated nonlinear transformati
ons. Unlike KCCA, DCCA does not require an inner product, and has the advantages
 of a parametric method: training time scales well with data size and the traini
ng data need not be referenced when computing the representations of unseen inst
ances. In experiments on two real-world datasets, we find that DCCA learns repr
esentations with significantly higher correlation than those learned by CCA and
KCCA. We also introduce a novel non-saturating sigmoid function based on the cub
e root that may be useful more generally in feedforward neural networks.
****
Consistency of Online Random Forests
Misha Denil, David Matheson, Nando Freitas
As a testament to their success, the theory of random forests has long been outp
aced by their application in practice. In this paper, we take a step towards nar

rowing this gap by providing a consistency result for online random forests.
****
Sparse Gaussian Conditional Random Fields: Algorithms, Theory, and Application t
o Energy Forecasting
Matt Wytock, Zico Kolter
This paper considers the sparse Gaussian conditional random field, a discriminat
ive extension of sparse inverse covariance estimation, where we use convex metho
ds to learn a high-dimensional conditional distribution of outputs given inputs.
 The model has been proposed by multiple researchers within the past year, yet p
revious papers have been substantially limited in their analysis of the method a
nd in the ability to solve large-scale problems. In this paper, we make three c
ontributions: 1) we develop a second-order active-set method which is several or
ders of magnitude faster that previously proposed optimization approaches for th
is problem 2) we analyze the model from a theoretical standpoint, improving upon
 past bounds with convergence rates that depend logarithmically on the data dime
nsion, and 3) we apply the method to large-scale energy forecasting problems, de
monstrating state-of-the-art performance on two real-world tasks.
****
Fast Image Tagging
Minmin Chen, Alice Zheng, Kilian Weinberger
Automatic image annotation is a difficult and highly relevant machine learning t
ask. Recent advances have significantly improved the state-of-the-art in retriev
al accuracy with algorithms based on nearest neighbor classification in carefull
y learned metric spaces. But this comes at a price of increased computational co
mplexity during training and testing. We propose FastTag, a novel algorithm that
 achieves comparable results with two simple linear mappings that are co-regular
ized in a joint convex loss function. The loss function can be efficiently optim
ized in closed form updates, which allows us to incorporate a large number of im
age descriptors cheaply. On several standard real-world benchmark data sets, we
demonstrate that FastTag matches the current state-of-the-art in tagging quality
, yet reduces the training and testing times by several orders of magnitude and
has lower asymptotic complexity.
****
Expensive Function Optimization with Stochastic Binary Outcomes
Matthew Tesch, Jeff Schneider, Howie Choset
Real world systems often have parameterized controllers which can be tuned to im
prove performance. Bayesian optimization methods provide for efficient optimizat
ion of these controllers, so as to reduce the number of required experiments on
the expensive physical system. In this paper we address Bayesian optimization in
 the setting where performance is only observed through a stochastic binary outc
ome – success or failure of the experiment. Unlike bandit problems, the goal is
to maximize the system performance after this offline training phase rather than
 minimize regret during training. In this work we define the stochastic binary
optimization problem and propose an approach using an adaptation of Gaussian Pro
cesses for classification that presents a Bayesian optimization framework for th
is problem. We propose an experiment selection metric for this setting based on
 expected improvement. We demonstrate the algorithm’s performance on synthetic
problems and on a real snake robot learning to move over an obstacle.
****
Multiple-source cross-validation
Krzysztof Geras, Charles Sutton
Cross-validation is an essential tool in machine learning and statistics. The ty
pical procedure, in which data points are randomly assigned to one of the test s
ets, makes an implicit assumption that the data are exchangeable. A common case
in which this does not hold is when the data come from multiple sources, in the
sense used in transfer learning. In this case it is common to arrange the cross-
validation procedure in a way that takes the source structure into account. Alth
ough common in practice, this procedure does not appear to have been theoretical
ly analysed. We present new estimators of the variance of the cross-validation,
both in the multiple-source setting and in the standard iid setting. These new e

stimators allow for much more accurate confidence intervals and hypothesis tests to compare algorithms.
****
Learning Triggering Kernels for Multi-dimensional Hawkes Processes
Ke Zhou, Hongyuan Zha, Le Song
How does the activity of one person affect that of another person? Does the strength of influence remain periodic or decay exponentially over time? In this paper, we study these critical questions in social network analysis quantitatively under the framework of multi-dimensional Hawkes processes. In particular, we focus on the nonparametric learning of the triggering kernels, and propose an algorithm \sf MMEL that combines the idea of decoupling the parameters through constructing a tight upper-bound of the objective function and application of Euler-Lagrange equations for optimization in infinite dimensional functional space. We show that the proposed method performs significantly better than alternatives in experiments on both synthetic and real world datasets.
****
On the difficulty of training recurrent neural networks
Razvan Pascanu, Tomas Mikolov, Yoshua Bengio
There are two widely known issues with properly training recurrent neural networks, the vanishing and the exploding gradient problems detailed in Bengio et al. (1994). In this paper we attempt to improve the understanding of the underlying issues by exploring these problems from an analytical, a geometric and a dynamical systems perspective. Our analysis is used to justify a simple yet effective solution. We propose a gradient norm clipping strategy to deal with exploding gradients and a soft constraint for the vanishing gradients problem. We validate empirically our hypothesis and proposed solutions in the experimental section.
****
Maxout Networks
Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, Yoshua Bengio
We consider the problem of designing models to leverage a recently introduced approximate model averaging technique called dropout. We define a simple new model called maxout (so named because its output is the max of a set of inputs, and because it is a natural companion to dropout) designed to both facilitate optimization by dropout and improve the accuracy of dropout's fast approximate model averaging technique. We empirically verify that the model successfully accomplishes both of these tasks. We use maxout and dropout to demonstrate state of the art classification performance on four benchmark datasets: MNIST, CIFAR-10, CIFAR-100, and SVHN.
****
Predictable Dual-View Hashing
Mohammad Rastegari, Jonghyun Choi, Shobeir Fakhraei, Daume Hal, Larry Davis
We propose a Predictable Dual-View Hashing (PDH) algorithm which embeds proximity of data samples in the original spaces. We create a cross-view hamming space with the ability to compare information from previously incomparable domains with a notion of ‘predictability’. By performing comparative experimental analysis on two large datasets, PASCAL-Sentence and SUN-Attribute, we demonstrate the superiority of our method to the state-of-the-art dual-view binary code learning algorithms.
****
Deep learning with COTS HPC systems
Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, Ng Andrew
Scaling up deep learning algorithms has been shown to lead to increased performance in benchmark tasks and to enable discovery of complex high-level features. Recent efforts to train extremely large networks (with over 1 billion parameters) have relied on cloud-like computing infrastructure and thousands of CPU cores. In this paper, we present technical details and results from our own system based on Commodity Off-The-Shelf High Performance Computing (COTS HPC) technology: a cluster of GPU servers with Infiniband interconnects and MPI. Our system is able to train 1 billion parameter networks on just 3 machines in a couple of days, and we show that it can scale to networks with over 11 billion parameters usi

ng just 16 machines. As this infrastructure is much more easily marshaled by others, the approach enables much wider-spread research with extremely large neural networks.
****
Nonparametric Mixture of Gaussian Processes with Constraints
James Ross, Jennifer Dy
Motivated by the need to identify new and clinically relevant categories of lung disease, we propose a novel clustering with constraints method using a Dirichlet process mixture of Gaussian processes in a variational Bayesian nonparametric framework. We claim that individuals should be grouped according to biological and/or genetic similarity regardless of their level of disease severity; therefore, we introduce a new way of looking at subtyping/clustering by recasting it in terms of discovering associations of individuals to disease trajectories (i.e., grouping individuals based on their similarity in response to environmental and/or disease causing variables). The nonparametric nature of our algorithm allows for learning the unknown number of meaningful trajectories. Additionally, we acknowledge the usefulness of expert guidance by providing for their input using must-link and cannot- link constraints. These constraints are encoded with Markov random fields. We also provide an efficient variational approach for performing inference on our model.
****
Scale Invariant Conditional Dependence Measures
Sashank J Reddi, Barnabas Poczos
In this paper we develop new dependence and conditional dependence measures and provide their estimators. An attractive property of these measures and estimators is that they are invariant to any monotone increasing transformations of the random variables, which is important in many applications including feature selection. Under certain conditions we show the consistency of these estimators, derive upper bounds on their convergence rates, and show that the estimators do not suffer from the curse of dimensionality. However, when the conditions are less restrictive, we derive a lower bound which proves that in the worst case the convergence can be arbitrarily slow similarly to some other estimators. Numerical illustrations demonstrate the applicability of our method.
****
Learning Policies for Contextual Submodular Prediction
Stephane Ross, Jiaji Zhou, Yisong Yue, Debadeepta Dey, Drew Bagnell
Many prediction domains, such as ad placement, recommendation, trajectory prediction, and document summarization, require predicting a set or list of options. Such lists are often evaluated using submodular reward functions that measure both quality and diversity. We propose a simple, efficient, and provably near-optimal approach to optimizing such prediction problems based on no-regret learning. Our method leverages a surprising result from online submodular optimization: a single no-regret online learner can compete with an optimal sequence of predictions. Compared to previous work, which either learn a sequence of classifiers or rely on stronger assumptions such as realizability, we ensure both data-efficiency as well as performance guarantees in the fully agnostic setting. Experiments validate the efficiency and applicability of the approach on a wide range of problems including manipulator trajectory optimization, news recommendation and document summarization.
****
Manifold Preserving Hierarchical Topic Models for Quantization and Approximation
Minje Kim, Paris Smaragdis
We present two complementary topic models to address the analysis of mixture data lying on manifolds. First, we propose a quantization method with an additional mid-layer latent variable, which selects only data points that best preserve the manifold structure of the input data. In order to address the case of modeling all the in-between parts of that manifold using this reduced representation of the input, we introduce a new model that provides a manifold-aware interpolation method. We demonstrate the advantages of these models with experiments on the hand-written digit recognition and the speech source separation tasks.

****
Safe Screening of Non-Support Vectors in Pathwise SVM Computation
Kohei Ogawa, Yoshiki Suzuki, Ichiro Takeuchi
In this paper, we claim that some of the non-support vectors (non-SVs) that have no influence on the classifier can be screened out prior to the training phase in pathwise SVM computation scenario, in which one is asked to train a sequence (or path) of SVM classifiers for different regularization parameters. Based on a recently proposed framework so-called safe screening rule, we derive a rule for screening out non-SVs in advance, and discuss how we can exploit the advantage of the rule in pathwise SVM computation scenario. Experiments indicate that our approach often substantially reduce the total pathwise computation cost.
****
Cost-sensitive Multiclass Classification Risk Bounds
Bernardo Ávila Pires, Csaba Szepesvari, Mohammad Ghavamzadeh
A commonly used approach to multiclass classification is to replace the 0-1 loss with a convex surrogate so as to make empirical risk minimization computationally tractable. Previous work has uncovered sufficient and necessary conditions for the consistency of the resulting procedures. In this paper, we strengthen these results by showing how the 0-1 excess loss of a predictor can be upper bounded as a function of the excess loss of the predictor measured using the convex surrogate. The bound is developed for the case of cost-sensitive multiclass classification and a convex surrogate loss that goes back to the work of Lee, Lin and Wahba. The bounds are as easy to calculate as in binary classification. Furthermore, we also show that our analysis extends to the analysis of the recently introduced “Simplex Coding” scheme.
****
Semi-supervised Clustering by Input Pattern Assisted Pairwise Similarity Matrix Completion
Jinfeng Yi, Lijun Zhang, Rong Jin, Qi Qian, Anil Jain
Many semi-supervised clustering algorithms have been proposed to improve the clustering accuracy by effectively exploring the available side information that is usually in the form of pairwise constraints. Despite the progress, there are two main shortcomings of the existing semi-supervised clustering algorithms. First, they have to deal with non-convex optimization problems, leading to clustering results that are sensitive to the initialization. Second, none of these algorithms is equipped with theoretical guarantee regarding the clustering performance. We address these limitations by developing a framework for semi-supervised clustering based on \it input pattern assisted matrix completion. The key idea is to cast clustering into a matrix completion problem, and solve it efficiently by exploiting the correlation between input patterns and cluster assignments. Our analysis shows that under appropriate conditions, only O(\log n) pairwise constraints are needed to accurately recover the true cluster partition. We verify the effectiveness of the proposed algorithm by comparing it to the state-of-the-art semi-supervised clustering algorithms on several benchmark datasets.
****
Learning the beta-Divergence in Tweedie Compound Poisson Matrix Factorization Models
Umut Simsekli, Ali Taylan Cemgil, Yusuf Kenan Yilmaz
In this study, we derive algorithms for estimating mixed β-divergences. Such cost functions are useful for Nonnegative Matrix and Tensor Factorization models with a compound Poisson observation model. Compound Poisson is a particular Tweedie model, an important special case of exponential dispersion models characterized by the fact that the variance is proportional to a power function of the mean. There are several well known matrix and tensor factorization algorithms that minimize the β-divergence; these estimate the mean parameter. The probabilistic interpretation gives us more flexibility and robustness by providing us additional tunable parameters such as power and dispersion. Estimation of the power parameter is useful for choosing a suitable divergence and estimation of dispersion is useful for data driven regularization and weighting in collective/coupled factorization of heterogeneous datasets. We present three inference algorithms for bo

th estimating the factors and the additional parameters of the compound Poisson distribution. The methods are evaluated on two applications: modeling symbolic r epresentations for polyphonic music and lyric prediction from audio features. Ou r conclusion is that the compound poisson based factorization models can be usef ul for sparse positive data.
****
Fast algorithms for sparse principal component analysis based on Rayleigh quotie nt iteration
Volodymyr Kuleshov
We introduce new algorithms for sparse principal component analysis (sPCA), a va riation of PCA which aims to represent data in a sparse low-dimensional basis. O ur algorithms possess a cubic rate of convergence and can compute principal comp onents with k non-zero elements at a cost of O(nk + k^3) flops per iteration. We observe in numerical experiments that these components are of equal or greater quality than ones obtained from current state-of-the-art techniques, but require between one and two orders of magnitude fewer flops to be computed. Conceptuall y, our approach generalizes the Rayleigh quotient iteration algorithm for comput ing eigenvectors, and can be interpreted as a type of second-order optimization method. We demonstrate the applicability of our algorithms on several datasets, including the STL-10 machine vision dataset and gene expression data.
****
Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Modeling
Amr Ahmed, Liangjie Hong, Alexander Smola
Much natural data is hierarchical in nature. Moreover, this hierarchy is often shared between different instances. We introduce the nested Chinese Restaurant Franchise Process as a means to obtain both hierarchical tree-structured repres entations for objects, akin to (but more general than) the nested Chinese Restau rant Process while sharing their structure akin to the Hierarchical Dirichlet P rocess. Moreover, by decoupling the \emphstructure generating part of the p rocess from the components responsible for the observations, we are able to app ly the same statistical approach to a variety of user generated data. In partic ular, we model the joint distribution of microblogs and locations for Twitter f or users. This leads to a 40% reduction in location uncertainty relative to the best previously published results. Moreover, we model documents from the NIPS papers dataset, obtaining excellent perplexity relative to (hierarchical) Pach inko allocation and LDA.
****
Tree-Independent Dual-Tree Algorithms
Ryan Curtin, William March, Parikshit Ram, David Anderson, Alexander Gray, Charl es Isbell
Dual-tree algorithms are a widely used class of branch-and-bound algorithms. Un fortunately, developing dual-tree algorithms for use with different trees and pr oblems is often complex and burdensome. We introduce a four-part logical split: the tree, the traversal, the point-to-point base case, and the pruning rule. W e provide a meta-algorithm which allows development of dual-tree algorithms in a tree-independent manner and easy extension to entirely new types of trees. Rep resentations are provided for five common algorithms; for k-nearest neighbor se arch, this leads to a novel, tighter pruning bound. The meta-algorithm also allo ws straightforward extensions to massively parallel settings.
****
Multilinear Multitask Learning
Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, Massimiliano Pont il
Many real world datasets occur or can be arranged into multi-modal structures. With such datasets, the tasks to be learnt can be referenced by multiple indices . Current multitask learning frameworks are not designed to account for the pre servation of this information. We propose the use of multilinear algebra as a n atural way to model such a set of related tasks. We present two learning methods ; one is an adapted convex relaxation method used in the context of tensor com

pletion. The second method is based on the Tucker decomposition and on alternating minimization. Experiments on synthetic and real data indicate that the multilinear approaches provide a significant improvement over other multitask learning methods. Overall our second approach yields the best performance in all data sets.
****
Online Learning under Delayed Feedback
Pooria Joulani, Andras Gyorgy, Csaba Szepesvari
Online learning with delayed feedback has received increasing attention recently due to its several applications in distributed, web-based learning problems. In this paper we provide a systematic study of the topic, and analyze the effect of delay on the regret of online learning algorithms. Somewhat surprisingly, it turns out that delay increases the regret in a multiplicative way in adversarial problems, and in an additive way in stochastic problems. We give meta-algorithms that transform, in a black-box fashion, algorithms developed for the non-delayed case into ones that can handle the presence of delays in the feedback loop. Modifications of the well-known UCB algorithm are also developed for the bandit problem with delayed feedback, with the advantage over the meta-algorithms that they can be implemented with lower complexity.
****
Adaptive Hamiltonian and Riemann Manifold Monte Carlo
Ziyu Wang, Shakir Mohamed, Nando Freitas
In this paper we address the widely-experienced difficulty in tuning Hamiltonian-based Monte Carlo samplers. We develop an algorithm that allows for the adaptation of Hamiltonian and Riemann manifold Hamiltonian Monte Carlo samplers using Bayesian optimization that allows for infinite adaptation of the parameters of these samplers. We show that the resulting sampling algorithms are ergodic, and demonstrate on several models and data sets that the use of our adaptive algorithms makes it is easy to obtain more efficient samplers, in some precluding the need for more complex models. Hamiltonian-based Monte Carlo samplers are widely known to be an excellent choice of MCMC method, and we aim with this paper to remove a key obstacle towards the more widespread use of these samplers in practice.
****
Coco-Q: Learning in Stochastic Games with Side Payments
Eric Sodomka, Elizabeth Hilliard, Michael Littman, Amy Greenwald
Coco (""cooperative/competitive"") values are a solution concept for two-player normal-form games with transferable utility, when binding agreements and side payments between players are possible. In this paper, we show that coco values can also be defined for stochastic games and can be learned using a simple variant of Q-learning that is provably convergent. We provide a set of examples showing how the strategies learned by the Coco-Q algorithm relate to those learned by existing multiagent Q-learning algorithms.
****
On A Nonlinear Generalization of Sparse Coding and Dictionary Learning
Jeffrey Ho, Yuchen Xie, Baba Vemuri
Existing dictionary learning algorithms are based on the assumption that the data are vectors in an Euclidean vector space, and the dictionary is learned from the training data using the vector space structure and its Euclidean metric. However, in many applications, features and data often originated from a Riemannian manifold that does not support a global linear (vector space) structure. Furthermore, the extrinsic viewpoint of existing dictionary learning algorithms becomes inappropriate for modeling and incorporating the intrinsic geometry of the manifold that is potentially important and critical to the application. This paper proposes a novel framework for sparse coding and dictionary learning for data on a Riemannian manifold, and it shows that the existing sparse coding and dictionary learning methods can be considered as special (Euclidean) cases of the more general framework proposed here. We show that both the dictionary and sparse coding can be effectively computed for several important classes of Riemannian manifolds, and we validate the proposed method using two well-known classification problems in computer vision and medical imaging analysis.

****

Estimation of Causal Peer Influence Effects

Panos Toulis, Edward Kao

The broad adoption of social media has generated interest in leveraging peer influence for inducing desired user behavior. Quantifying the causal effect of peer influence presents technical challenges, however, including how to deal with social interference, complex response functions and network uncertainty. In this paper, we extend potential outcomes to allow for interference, we introduce well-defined causal estimands of peer-influence, and we develop two estimation procedures: a frequentist procedure relying on a sequential randomization design that requires knowledge of the network but operates under complicated response functions, and a Bayesian procedure which accounts for network uncertainty but relies on a linear response assumption to increase estimation precision. Our results show the advantages and disadvantages of the proposed methods in a number of situations.

****