27th International Conference on Machine Learning

(ICML 2010)

Haifa, Israel
21 – 24 June 2010

Volume 1 of 2

ISBN: 978-1-63266-059-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Printed by Curran Associates, Inc. (2014)

International Machine Learning Society (IMLS)
c/o Priscilla Rasmussen
209 North Eighth Street
Stroudsburg PA 18360

Phone: (570) 476-8006
Fax: (570) 476-0860

rasmussen@ptd.net

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Awards: Best Paper Awards (Oren, Wed, 8.30; Wed 14.30; Thu, 8.30)

Hilbert Space Embeddings of Hidden Markov Models (1)
Hidden Markov Models (HMMs) are important tools for modeling sequencedata. However, they are restricted to discrete latent states, and arelargely restricted to Gaussian and discrete observations. And,learning algorithms for HMMs have predominantly relied on local searchheuristics, with the exception of spectral methods such as thosedescribed below. We propose a nonparametric HMM thatextends traditional HMMs to structured and non-Gaussian continuous distributions. Furthermore, we derive a local-minimum-free kernel spectral algorithm for learning these HMMs. We apply our method to robot vision data, slot car inertialsensor data and audio event classification data, and show that inthese applications, embedded HMMs exceed the previous state-of-the-artperformance.
L. Song, B. Boots, S. Saddiqi, G. Gordon, A. Smola

On the Consistency of Ranking Algorithms (9)
We present a theoretical analysis of supervised ranking, providing necessary and sufficient conditions for the asymptotic consistency of algorithms based on minimizing a surrogate loss function. We show that many commonly used surrogate losses are inconsistent; surprisingly, we show inconsistency even in low-noise settings. We present a new value-regularized linear loss, establish its consistency under reasonable assumptions on noise, and show that it outperforms conventional ranking losses in a collaborative filtering experiment.
J. Duchi, L. Mackey, M. Jordan

Modeling Transfer Learning in Human Categorization with the Hierarchical Dirichlet Process (17)
Transfer learning can be described as the distillation of abstract knowledge from one learning domain or task and the reuse of that knowledge in a related domain or task. In categorization settings, transfer learning is the modification by past experience of prior expectations about what types of categories are likely to exist in the world. While transfer learning is an important and active research topic in machine learning, there have been few studies of transfer learning in human categorization. We propose an explanation for transfer learning effects in human categorization, implementing a model from the statistical machine learning literature -- the hierarchical Dirichlet process (HDP) -- to make empirical evaluations of its ability to explain these effects. We present two laboratory experiments which measure the degree to which people engage in transfer learning in a controlled setting, and we compare our model to their performance. We find that the HDP provides a good explanation for transfer learning exhibited by human learners.
K. Canini, T. Griffiths

1a: Topic Models and Matrix Factorization (Alon, Tue 10.30)

Spherical Topic Models (25)
We introduce the Spherical Admixture Model (SAM), a Bayesian topic model for arbitrary L2 normalized data. SAM maintains the same hierarchical structure as Latent Dirichlet Allocation (LDA), but models documents as points on a high-dimensional spherical manifold, allowing a natural likelihood parameterization in terms of cosine distance. Furthermore, SAM topics are capable of assigning negative weight to terms and can model word absence/presence unlike previous models. Performance is evaluated empirically both subjectively as a topic model using human raters and across several disparate classification tasks, from natural language processing and computer vision.
J. Reisinger, A. Waters, B. Silverthorn, R. Mooney

Conditional Topic Random Fields (33)
Generative topic models such as LDA are limited by their inability to utilize nontrivial input features to enhance their performance, and many topic models assume that topic assignments of different words are conditionally independent. Some work exists to address the second limitation but no work exists to address both. This paper presents a conditional topic random field (CTRF) model, which can use arbitrary non local features about words and documents and incorporate the Markov dependency between topic assignments of neighboring words. We develop an efficient variational inference algorithm that scales linearly in terms of topic numbers, and a maximum likelihood estimation (MLE) procedure for parameter estimation. For the supervised version of CTRF, we also develop an arguably more discriminative max-margin learning method. We evaluate CTRF on real review rating data and demonstrate the advantages of CTRF over generative competitors, and we show the advantages of max-margin learning over MLE.
J. Zhu, E. Xing

A Language-based Approach to Measuring Scholarly Impact (41)

Identifying the most influential documents in a corpus is an important problem in many fields, from information science and historiography to text summarization and news aggregation. Unfortunately, traditional bibliometrics such as citations are often not available. We propose using changes in the thematic content of documents over time to measure the importance of individual documents within the collection. We describe a dynamic topic model for both quantifying and qualifying the impact of these documents. We validate the model by analyzing three large corpora of scientific articles. Our measurement of a document's impact correlates significantly with its number of citations.
S. Gerrish, D. Blei

Mixed Membership Matrix Factorization (49)
Discrete mixed membership modeling and continuous latent factor modeling (also known as matrix factorization) are two popular, complementary approaches to dyadic data analysis. In this work, we develop a fully Bayesian framework for integrating the two approaches into unified Mixed Membership Matrix Factorization (M3F) models. We introduce two M3F models, derive Gibbs sampling inference procedures, and validate our methods on the Each Movie, Movie Lens, and Netflix Prize collaborative filtering datasets. We find that, even when fitting fewer parameters, the M3F models outperform state-of-the-art latent factor approaches on all benchmarks, yielding the greatest gains in accuracy on sparsely-rated, high-variance items.
L. Mackey, D. Weiss, M. Jordan

1b: Reinforcement Learning 1 (Tamar, Tue 10.30)

Least-Squares lambda Policy Iteration: Bias-Variance Trade-off in Control Problems (57)
In the context of large space MDPs with linear value function approximation, we introduce a new approximate version of lambda-Policy Iteration (Berteskas and Ioffe, 1996), a method that generalizes Value Iteration and Policy Iteration. Our approach, called Least-Squares lambda Policy Iteration, generalizes LSPI (Lagoudakis & Parr, 2003) which makes efficient use of training samples compared to classical temporal-differences methods. The motivation of our work is to exploit the lambda parameter within the least-squares context, and without having to generate new sample sat each iteration or to know a model of the MDP. We provide a performance bound that shows the soundness of the algorithm. We show empirically on a simple chain problem and on the Tetris game that this lambda parameter acts as a bias-variance trade-off that may improve the convergence and the performance of the policy obtained.
C. Thiery, B. Scherrer

Finite-Sample Analysis of LSTD (65)
In this paper we consider the problem of policy evaluation in reinforcement learning, i.e., learning the value function of a fixed policy, using the least-squares temporal-difference (LSTD) learning algorithm. We report a finite-sample analysis of LSTD. We first derive a bound on the performance of the LSTD solution evaluated at the states generated by the Markov chain and used by the algorithm to learn an estimate of the value function. This result is general in the sense that no assumption is made on the existence of a stationary distribution for the Markov chain. We then derive generalization bounds in the case when the Markov chain possesses a stationary distribution and is β-mixing.
A. Lazaric, M. Ghavamzadeh, R. Munos

Convergence of Least Squares Temporal Difference Methods Under General Conditions (73)
We consider approximate policy evaluation for finite state and action Markov decision processes (MDP) in the off-policy learning context and with the simulation-based least squares temporal difference algorithm, LSTD(λ). We establish for the discounted cost criterion that the off-policy LSTD(λ) converges almost surely under mild, minimal conditions. We also analyze other convergence and boundedness properties of the iterates involved in the algorithm, and based on them, we suggest a modification in its practical implementation. Our analysis uses theories of both finite space Markov chains and Markov chains on topological spaces.
H. Yu

Should one compute the Temporal Difference fix point or minimize the Bellman Residual? The unified oblique projection view (81)

We investigate projection methods, for evaluating a linear approximation of the value function of a policy in a Markov Decision Process context. We consider two popular approaches, the one-step Temporal Difference fix-point computation (TD(0)) and the Bellman Residual (BR) minimization. We describe examples, where each method outperforms the other. We highlight a simple relation between the objective function they minimize, and show that while BR enjoys a performance guarantee, TD(0) does not in general. We then propose a unified view in terms of oblique projections of the Bellman equation, which substantially simplifies and extends the characterization of Schoknecht (2002) and the recent analysis of Yu and Bertsekas (2008). Eventually, we describe some simulations that suggest that if the TD(0) solution is usually slightly better than the BR solution, its inherent numerical instability makes it very bad in some cases, and thus worse on average.
B. Scherrer

1c: Ensemble Methods (Rimon, Tue 10.30)

Fast boosting using adversarial bandits (89)
In this paper we apply multi-armed bandits (MABs) to improve the computational complexity of AdaBoost. AdaBoost constructs a strong classifier in a stepwise fashion by selecting simple base classifiers and using their weighted ``vote'' to determine the final classification. We model this stepwise base classifier selection as a sequential decision problem, and optimize it with MABs where each arm represents a subset of the base classifier set. The MAB gradually learns the ``usefulness'' of the subsets, and selects one of the subsets in each iteration. AdaBoost then searches only this subset instead of optimizing the base classifier over the whole space. The main improvement of this paper over a previous approach is that we use an adversarial bandit algorithm instead of stochastic bandits. This choice allows us to prove a weak-to-strong-learning theorem, which means that the proposed technique remains a boosting algorithm in a formal sense. We demonstrate on benchmark datasets that our technique can achieve a generalization performance similar to standard AdaBoost for a computational cost that is an order of magnitude smaller.
Robert Busa-Fekete, Balazs Kegl

Boosting for Regression Transfer (97)
The goal of transfer learning is to improve the learning of a new target concept given knowledge of related source concept(s). We introduce the first boosting-based algorithms for transfer learning that apply to regression tasks. First, we describe two existing classification transfer algorithms, ExpBoost and TrAdaBoost, and show how they can be modified for regression. We then introduce extensions of these algorithms that improve performance significantly on controlled experiments in a wide range of test domains.
D. Pardoe, P. Stone

Supervised Aggregation of Classifiers using Artificial Prediction Markets (105)
Prediction markets are used in real life to predict outcomes of interest such as presidential elections. In this work we introduce a mathematical theory for Artificial Prediction Markets for supervised classifier aggregation and probability estimation. We introduce the artificial prediction market as a novel way to aggregate classifiers. We derive the market equations to enforce total budget conservation, show the market price uniqueness and give efficient algorithms for computing it. We show how to train the market participants by updating their budgets using training examples. We introduce classifier specialization as a new differentiating characteristic between classifiers. Finally, we present experiments using random decision rules as specialized classifiers and show that the prediction market consistently outperforms Random Forest on real and synthetic data of varying degrees of difficulty.
N. Lay, A. Barbu

Boosting Classifiers with Tightened L0-Relaxation Penalties (113)
We propose a novel boosting algorithm which improves on current algorithms for weighted voting classification in striking a better balance between classification accuracy and sparsity of the weight vector. In order to justify our optimization formulations, we first consider a novel integer linear program as a model forsparse classifier selection, generalizing the minimum disagreementhalfspace problem, whose complexity has been investigated in computational learning theory. Specifically, our mixed integer problem is that of finding a separating hyperplane with minimum empirical error subject to an L0-norm penalty. We find that common soft margin linear programming formulations for robust classification are equivalent to a

continuous relaxation of our model. Since the initial continuous relaxation is weak, we suggesta tighter relaxation, using novel cutting planes, to better approximate the integer solution. To solve this relaxation, we propose a new boosting algorithm based on linear programming with dynamic generation of variables and constraints. We demonstrate the classification performance of our proposed algorithm with experimental results, and justify our selection of parameters using a minimum description length, compression intrepretation of learning.
N. Goldberg, J. Eckstein

1d: Statistical Relational Learning (Hadas, Tue 10.30)

Probabilistic Backward and Forward Reasoning in Stochastic Relational Worlds (121)
Inference in graphical models has emerged as a promising technique for planning. A recent approach to decision-theoretic planning in relational domains uses forward inference in dynamic Bayesian networks compiled from learned probabilistic relational rules. Inspired by work in non-relational domains with small state spaces, we derive a backpropagation method for such nets in relational domains starting from a goal state mixture distribution. We combine this with forward reasoning in a bidirectional two-filter approach. We perform experiments in a complex 3D simulated desktop environment with an articulated manipulator and realistic physics. Empirical results show that bidirectional probabilistic reasoning can lead to more efficient and accurate planning in comparison to pure forward reasoning.
T. Lang, M. Toussaint

Active Learning for Networked Data (129)
We introduce a novel active learning algorithm for classification of network data. In this setting, training instances are connected by a set of links to form a network, the labels of linked nodes are correlated, and the goal is to exploit these dependencies and accurately label the nodes. This problem arises in many domains, including social and biological network analysis and document classification, and there has been much recent interest in methods that collectively classify the nodes in the network. While in many cases labeled examples are expensive, often network information is available. We show how an active learning algorithm can take advantage of network structure. Our algorithm effectively exploits the links between instances and the interaction between the local and collective aspects of a classifier to improve the accuracy of learning from fewer labeled examples. We experiment with two real-world benchmark collective classification domains, and show that we are able to achieve extremely accurate results even when only a small fraction of the data is labeled.
M. Bilgic, L. Mihalkova, L. Getoor

Learning Markov Logic Networks Using Structural Motifs (137)
Markov logic networks (MLNs) use first-order formulas to define features of Markov networks. Current MLN structure learners can only learn short clauses (4-5 literals) due to extreme computational costs, and thus are unable to represent complex regularities in data. To address this problem,we present LSM, the first MLN structure learner capable of efficiently and accurately learning long clauses. LSM is based on the observation that relational data typically contains patterns that are variations of the same structural motifs. By constraining the search for clauses to occur within motifs, LSM can greatly speed up the search and thereby reduce the cost of finding long clauses. LSM uses random walks to identify densely connected objects in data, and groups them and their associated relations into a motif.Our experiments on three real-world datasets show that our approach is 2-5 orders of magnitude faster than the state-of-the-art ones, while achieving the same or better predictive performance.
S. Kok, P. Domingos

Bottom-Up Learning of Markov Network Structure (145)
The structure of a Markov network is typically learned using top-down search. At each step, the search specializes a feature by conjoining it to the variable or feature that most improves the score. This is inefficient, testing many feature variations with no support in the data, and highly prone to local optima. We propose bottom-up search as an alternative, inspired by the analogous approach in the field of rule induction. Our BLM algorithm starts with each complete training example as a long feature, and repeatedly generalizes a feature to match its k nearest examples by dropping variables. An extensive empirical evaluation demonstrates that BLM is both faster and more accurate than the standard top-down approach,

and also outperforms other state-of-the-art methods.
J. Davis, P. Domingos

1e: Large-Scale Learning and Optimization (Arava, Tue 10.30)

A fast natural Newton method (153)
Nowadays, for many tasks such as object recognition or language modeling, data is plentiful. As such, an important challenge has become to find learning algorithms which can make use of all the available data. In this setting, called ``large-scale learning'' by Bottou and Bousquet (2008), learning and optimization become different and powerful optimization algorithms are suboptimal learning algorithms. While most efforts are focused on adapting optimization algorithms for learning by efficiently using the information contained in the Hessian, Le Roux et al. (2008) exploited the special structure of the learning problem to achieve faster convergence. In this paper, we investigate a natural way of combining these two directions to yield fast and robust learning algorithms.
N. Le Roux, A. Fitzgibbon

A scalable trust-region algorithm with application to mixed-norm regression (161)
We present a new algorithm for minimizing a convex loss-function subject to regularization. Our framework applies to numerous problems in machine learning and statistics; notably, for sparsity-promoting regularizers such as l_1 or $l_{1,\infty}$ norms it enables efficient computation of sparse solutions. Our approach is based on the trust-region framework with non smooth objectives, which allows us to build on known results to provide convergence analysis. We avoid the computational overheads associated with the conventional Hessian approximation used by trust-region methods by instead using a simple separable quadratic approximation. This approximation also enables use of proximity operators for tackling non smooth regularizers. We illustrate the versatility of our resulting algorithm by specializing it to three mixed-norm regression problems: group lasso [36], group logistic regression [21], and multi-task lasso [19]. We experiment with both synthetic and real-world large-scale data—our method is seen to be competitive, robust, and scalable.
D. Kim, S. Sra, I. Dhillon

Making Large-Scale Nystrom Approximation Possible (169)
The Nystrom method is an efficient technique for the eigen value decomposition of large kernel matrices. However, in order to ensure an accurate approximation, a sufficiently large number of columns have to be sampled. On very large data sets, the SVD step on the resultant data sub matrix will soon dominate the computations and become prohibitive. In this paper, we propose an accurate and scalable Nystrom scheme that first samples a large column subset from the input matrix, but then only performs an approximate SVD on the inner sub matrix by using the recent randomized low-rank matrix approximation algorithms. Theoretical analysis shows that the proposed algorithm is as accurate as the standard Nystrom method that directly performs a large SVD on the inner sub matrix On the other hand, its time complexity is only as low as performing a small SVD. Experiments are performed on a number of large-scale data sets for low-rank approximation and spectral embedding. In particular, spectral embedding of a MNIST data set with 3.3 million examples takes less than an hour on a standard PC with 4G memory.
M. Li, J. Kwok, B.-L. Lu

Gaussian Covariance and Scalable Variational Inference (177)
We analyze computational aspects of variational approximate inference techniques for sparse linear models, which have to be understood to allow for large scale applications. Gaussian covariances play a key role,whose approximation is computationally hard. While most previous methods gain scalability by not even representing most posterior dependencies, harmful factorization assumptions can be avoided by employing data-dependent low-rank approximations instead. We provide theoretical and empirical insights into algorithmic and statistical consequences of low-rank covariance approximation errors on decision outcomes in nonlinear sequential Bayesian experimental design.
M. Seeger

2a: Matrix Factorization and Recommendation (Alon, Tue

13.30)

Implicit Regularization in Variational Bayesian Matrix Factorization (185)
Matrix factorization into the product of low-rank matrices induces non-identifiability, i.e.,the mapping between the target matrix and factorized matrices is not one-to-one.In this paper, we theoretically investigate the influence of non-identifiability on Bayesian matrix factorization.More specifically, we show that a variational Bayesian method involves regularization effect even when the prior is non-informative,which is intrinsically different from the maximum a posteriori approach.We also extend our analysis to empirical Bayes scenarios where hyper parameters are also learned from data.
S. Nakajima, M. Sugiyama

Bayesian Nonparametric Matrix Factorization for Recorded Music (193)
Recent research in machine learning has focused on breaking audio spectrograms into separate sources of sound using latent variable decompositions. These methods require that the number of sources be specified in advance, which is not always possible. To address this problem, we develop Gamma Process Nonnegative Matrix Factorization (GaP-NMF), a Bayesian nonparametric approach to decomposing spectrograms. The assumptions behind GaP-NMF are based on research in signal processing regarding the expected distributions of spectrogram data, and GaP-NMF automatically discovers the number of latent sources. We derive a mean-field variational inference algorithm and evaluate GaP-NMF on both synthetic data and recorded music.
M. Hoffman, D. Blei, P. Cook

A Simple Algorithm for Nuclear Norm Regularized Problems (201)
Optimization problems with a nuclear norm regularization, such as e.g. low norm matrix factorizations, have seen many applications recently. We propose a new approximation algorithm building upon the recent sparse approximate SDP solver of (Hazan, 2008). The experimental efficiency of our method is demonstrated on large matrix completion problems such as the Netflix data set. The algorithm comes with strong convergence guarantees, and can be interpreted as a first theoretically justified variant of Simon-Funk-type SVD heuristics. The method is free of tuning parameters, and very easy to parallelize.
M. Jaggi, M. Sulovsky

Transfer Learning for Collective Link Prediction in Multiple Heterogenous Domains (209)
Link prediction is a key technique in many applications such as recommended systems, where potential links between users and items need to be predicted. A challenge in link prediction is the data sparsity problem. In this paper, we address this problem by jointly considering multiple heterogeneous link prediction tasks such as predicting links between users and different types of items including books, movies and songs, which we refer to as the collective link prediction (CLP) problem. We propose a nonparametric Bayesian framework for solving the CLP problem, which allows knowledge to be adaptively transferred across heterogeneous tasks while taking into account the similarities between tasks. We learn the inter-task similarity automatically. We also introduce link functions for different tasks to correct their biases and skewness of distributions in their link data. We conduct experiments on several real world datasets and demonstrate significant improvements over several existing state-of-the-art methods.
B. Cao, N. Liu, Q. Yang

2b: Reinforcement Learning 2 (Tamar, Tue 13.30)

Approximate Predictive Representations of Partially Observable Systems (217)
We provide a novel view of learning an approximate model of a partially observable environment from data and present a simple implementation of the idea. The learned model abstracts away unnecessary details of the agent's experience and focuses only on making certain predictions of interest. We illustrate our approach empirically in small computational examples, demonstrating the data efficiency of the algorithm.
D. Precup, M. Dinculescu

Constructing States for Reinforcement Learning (225)
POMDPs are the models of choice for reinforcement learning (RL) tasks where the environment cannot be

observed directly. In many applications we need to learn the POMDP structure and parameters from experience and this is considered to be a difficult problem. In this paper we address this issue by modeling the hidden environment with a novel class of models that are less expressive, but easier to learn and plan with than POMDPs. We call these models *deterministic Markov models* (DMMs), which are deterministic-probabilistic finite automata from learning theory, extended with actions to the sequential (rather than i.i.d.) setting. Conceptually, we extend the Utile Suffix Memory method of McCallum to handle long term memory. We describe DMMs, give Bayesian algorithms for learning and planning with them and also present experimental results for some standard POMDP tasks and tasks to illustrate its efficacy.
M. M. Mahmud

Temporal Difference Bayesian Model Averaging: A Bayesian Perspective on Adapting Lambda (233)
Temporal difference (TD) algorithms are attractive for reinforcement learning due to their ease-of-implementation and use of ``bootstrapped'' return estimates to make efficient use of sampled data. In particular, TD(lambda) methods comprise a family of reinforcement learning algorithms that often yield fast convergence by averaging multiple estimators of the expected return. However, TD(lambda) chooses a very specific way of averaging these estimators based on the fixed parameter lambda, which may not lead to optimal convergence rates in all settings. In this paper, we derive an automated Bayesian approach to setting lambda that we call temporal difference Bayesian model averaging (TD-BMA). Empirically, TD-BMA always performs as well and often much better than the best fixed lambda for TD(lambda) (even when performance for different values of lambda varies across problems) without requiring that lambda or any analogous parameter be manually tuned.
C. Downey, S. Sanner

Bayesian Multi-Task Reinforcement Learning (241)
We consider the problem of multi-task reinforcement learning where the learner is provided with a set of tasks, for which only a small number of samples can be generated for any given policy. As the number of samples may not be enough to learn an accurate evaluation of the policy, it would be necessary to identify classes of tasks with similar structure and to learn them jointly. We consider the case where the tasks share structure in their value functions, and model this by assuming that the value functions are all sampled from a common prior. We adopt the Gaussian process temporal-difference value function model and use a hierarchical Bayesian approach to model the distribution over the value functions. We study two cases, where all the value functions belong to the same class and where they belong to an undefined number of classes. For each case, we present a hierarchical Bayesian model, and derive inference algorithms for (i) joint learning of the value functions, and (ii) efficient transfer of the information gained in (i) to assist learning the value function of a newly observed task.
A. Lazaric, M. Ghavamzadeh

2c: Deep Learning 1 (Rimon, Tue 13.30)

3D Convolutional Neural Networks for Human Action Recognition (249)
We consider the fully automated recognition of actions in uncontrolled environment. Most existing work relies on domain knowledge to construct complex handcrafted features from inputs. In addition, the environments are usually assumed to be controlled. Convolutional neural networks (CNNs) are a type of deep models that can act directly on the raw inputs, thus automating the process of feature construction. However, such models are currently limited to handle 2D inputs. In this paper, we develop a novel 3D CNN model for action recognition. This model extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generates multiple channels of information from the input frames, and the final feature representation is obtained by combining information from all channels. We apply the developed model to recognize human actions in real-world environment, and it achieves superior performance without relying on handcrafted features.
S. Ji, W. Xu, M. Yang, K. Yu

Deep networks for robust visual recognition (257)
Deep Belief Networks (DBNs) are hierarchical generative models which have been used successfully to

model high dimensional visual data. However, they are not robust to common variations such as occlusion and random noise. We explore two strategies for improving the robustness of DBNs. First, we show that a DBN with sparse connections in the first layer is more robust to variations that are not in the training set. Second, we develop a probabilistic denoising algorithm to determine a subset of the hidden layer nodes to unclamp. We show that this can be applied to any feed forward network classifier with localized first layer connections. Recognition results after denoising are significantly better over the standard DBN implementations for various sources of noise.
Y. Tang, C. Eliasmith

Boosted Backpropagation Learning for Training Deep Modular Networks (265)
Divide-and-conquer is key to building sophisticated learning machines: hard problems are solved by composing a network of modules that solve simpler problems (LeCun et al., 1998; Rohde, 2002; Bradley, 2009). Many such existing systems rely on learning algorithms which are based on simple parametric gradient descent where the parametrization must be predetermined, or more specialized per-application algorithms which are usually ad-hoc and complicated. We present a novel approach for training generic modular networks that uses two existing techniques: the error propagation strategy of backpropagation and more recent research on descent in spaces of functions (Mason et al., 1999; Scholkopf & Smola, 2001). Combining these two methods of optimization gives a simple algorithm for training heterogeneous networks of functional modules using simple gradient propagation mechanics and established learning algorithms. The resulting separation of concerns between learning individual modules and error propagation mechanics eases implementation, enables a larger class of modular learning strategies, and allows per-module control of complexity/regularization. We derive and demonstrate this functional backpropagation and contrast it with traditional gradient descent in parameter space, observing that in our example domain the method is significantly more robust to local optima.
A. Grubb, D. Bagnell

Deep learning via Hessian-free optimization (273)
We develop a 2nd-order optimization method based on the ``Hessian-free approach, and apply it to training deep auto-encoders. Without using pre-training, we obtain results superior to those reported by Hinton & Salakhutdinov (2006) on the same tasks they considered. Our method is practical, easy to use, scales nicely to very large datasets, and isn't limited in applicability to auto-encoders, or any specific model class. We also discuss the issue of ``pathological curvature as a possible explanation for the difficulty of deep-learning and how 2nd-order optimization, and our method in particular, effectively deals with it.
J. Martens

2d: Multi-Task and Transfer Learning (Hadas, Tue 13.30)

Active Learning for Multi-Task Adaptive Filtering (281)
In this paper, we propose an Active Learning (AL) framework for the Multi-Task Adaptive Filtering (MTAF) problem. Specifically, we explore AL approaches to rapidly improve an MTAF system, based on Dirichlet Process priors, with minimal user/task-level feedback. The proposed AL approaches select instances for delivery with a two-fold objective: 1) Improve future task-specific system performance based on feedback received on delivered instances for that task, 2) Improve the future overall system performance, thereby benefiting other tasks in the system, based on feedback received on delivered instances for a particular task. Current AL approaches focus only on the first objective. For satisfying both goals, we define a new scoring function called Utility Gain to estimate the perceived improvements in task-specific and global models. In our experiments on standard benchmark datasets, we observed that global AL approaches that additionally take into account the potential benefit of feedback to other tasks in the system performed better than the task-specific approach that focused only the benefit of the current task.
A. Harpale, Y. Yang

Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity (289)
We consider the problem of learning a sparse multi-task regression,where the structure in the output scan be represented as a tree with leaf nodes as outputs and internal nodes as clusters of the outputs at multiple granularity. Our goal is to recover the common set of relevant inputs for each output cluster. Assuming that the tree structure is available as prior knowledge,we formulate this problem as a new multi-

task regularized regression called tree-guided group lasso. Our structured regularization is based on a group-lasso penalty, where groups are defined with respect to the tree structure. We describe a systematic weighting scheme for the groups in the penalty such that each output variable is penalized in a balanced manner even if the groups overlap.We present an efficient optimization method that can handle a large-scale problem.
S. Kim, E. Xing

OTL: A Framework of Online Transfer Learning (297)
In this paper, we investigate a new machine learning framework called ``Online Transfer Learning (OTL) that aims to transfer knowledge from some source domain to an online learning task on a target domain. We do not assume the target data follows the same class or generative distribution as the source data, and our key motivation is to improve a supervised online learning task in a target domain by exploiting the knowledge that had been learned from large amount of training data in source domains. OTL is in general challenging since data in both domains not only can be different in their class distributions but can be also different in their feature representations. As a first attempt to this problem, we propose techniques to address two kinds of OTL tasks: one is to perform OTL in a homogeneous domain, and the other is to perform OTL across heterogeneous domains. We show the mistake bounds of the proposed OTL algorithms, and empirically examine their performance on several challenging OTL tasks. Encouraging results validate the efficacy of our techniques.
P. Zhao, S. C.H. Hoi

2e: Ranking and Preference Learning (Arava, Tue 13.30)

Label Ranking under Ambiguous Supervision for Learning Semantic Correspondences (305)
This paper studies the problem of learning from ambiguous supervision, focusing on the task of learning semantic correspondences. A learning problem is said to be ambiguously supervised when, for a given training input, a set of output candidates is provided with no prior of which one is correct. We propose to tackle this problem by solving a related unambiguous task with a label ranking approach and show how and why this performs well on the original task, via the method of task-transfer. We apply it to learning to match natural language sentences to a structured representation of their meaning and empirically demonstrate that this competes with the state-of-the-art on two benchmarks.
A. Bordes, N. Usunier, J. Weston

Label Ranking Methods based on the Plackett-Luce Model (313)
This paper introduces two new methods for label ranking based on a probabilistic model of ranking data, called the Plackett-Luce model. The idea of the first method is to use the PL model to fit locally constant probability models in the context of instance-based learning. As opposed to this, the second method estimates a global model in which the PL parameters are represented as functions of the instance. Comparing our methods with previous approaches to label ranking, we find that they offer a number of advantages. Experimentally, we moreover show that they are highly competitive to start-of-the-art methods in terms of predictive accuracy, especially in the case of training data with incomplete ranking information.
W. Cheng, K. Dembczynski, E. Hullermeier

Metric Learning to Rank (321)
We study metric learning as problem of information retrieval. We present a general metric learning algorithm, based on the structural SVM framework, to learn a metric such that rankings of data induced by distance from a query can be optimized against various ranking measures, such as AUC, Precision-at-k, MRR, MAP or NDCG. We demonstrate experimental results on standard classification data sets, and a large-scale online dating recommendation problem.
B. McFee, G. Lanckriet

Learning Hierarchical Riffle Independent Groupings from Rankings (329)
Riffled independence is a generalized notion of probabilistic independence that has been shown to be naturally applicable to ranked data. In the riffled independence model, one assigns rankings to two disjoint sets of items independently, then in a second stage, interleaves (or riffles) the two rankings together to form a full ranking, as if by shuffling a deck of cards. Because of this interleaving stage, it is much more

difficult to detect riffled independence than ordinary independence. In this paper, we provide the first automated method for discovering sets of items which are riffle independent from a training set of rankings. We show that our clustering-like algorithms can be used to discover meaningful latent coalitions from real preference ranking datasets and to learn the structure of hierarchically decomposable models based on riffled independence.
J. Huang, C. Guestrin

3a: Latent-Variable Models (Alon, Tue 15.40)

The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling (337)
The hierarchical Dirichlet process (HDP) is a Bayesian nonparametric mixed membership model---each data point is modeled with a collection of components of different proportions. Though powerful, the HDP makes an assumption that the probability of a component being exhibited by a data point is positively correlated with its proportion within that data point. This might be an undesirable assumption. For example, in topic modeling, a topic (component) might be rare throughout the corpus but dominant within those documents (data points) where it occurs. We develop the IBP compound Dirichlet process (ICD), a Bayesian nonparametric prior that decouples across-data prevalence and within-data proportion in a mixed membership model. The ICD combines properties from the HDP and the Indian buffet process (IBP), a Bayesian nonparametric prior on binary matrices. The ICD assigns a subset of the shared mixture components to each data point. This subset, the data point's ``focus'', is determined independently from the amount that each of its components contribute. We develop an ICD mixture model for text, the focused topic model (FTM), and show superior performance over the HDP-based topic model.
S. Williamson, C. Wang, K. Heller, D. Blei

Forgetting Counts: Constant Memory Inference for a Dependent Hierarchical Pitman-Yor Process (345)
We propose a novel dependent hierarchical Pitman-Yor process model for discrete data. An incremental Monte Carlo inference procedure for this model is developed. We show that inference in this model can be performed in constant space and linear time. The model is demonstrated in a discrete sequence prediction task where it is shown to achieve state of the art sequence prediction performance while using significantly less memory.
N. Bartlett, D. Pfau, F. Wood

A Stick-Breaking Construction of the Beta Process (353)
We present and derive a new stick-breaking construction of the beta process. The construction is closely related to a special case of the stick-breaking construction of the Dirichlet process (Sethuraman, 1994) applied to the beta distribution. We derive an inference procedure that relies on Monte Carlo integration to reduce the number of parameters to be inferred, and present results on synthetic data, the MNIST handwritten digits data set and a time-evolving gene expression data set.
J. Paisley, A. Zaas, C. Woods, G. Ginsburg, L. Carin

Distance Dependent Chinese Restaurant Processes (361)
We develop the distance dependent Chinese restaurant process (CRP), a flexible class of distributions over partitions that allows for non-exchangeability. This class can be used to model dependencies between data in infinite clustering models, including dependencies across time or space. We examine the properties of the distance dependent CRP, discuss its connections to Bayesian nonparametric mixture models, and derive a Gibbs sampler for both observed and mixture settings. We study its performance with time-dependent models and three text corpora. We show that relaxing the assumption of exchangeability with distance dependent CRPs can provide a better fit to sequential data. We also show its alternative formulation of the traditional CRP leads to a faster-mixing Gibbs sampling algorithm than the one based on the original formulation.
D. Blei, P. Frazier

3b: Reinforcement Learning 3 (Tamar, Tue 15.40)

Generalizing Apprenticeship Learning across Hypothesis Classes (369)
This paper develops a generalized apprenticeship learning protocol for reinforcement-learning agents with access to a teacher who provides policy traces (transition and reward observations). We characterize sufficient conditions of the underlying models for efficient apprenticeship learning and link this criteria to two established learn ability classes (KWIK and Mistake Bound). We then construct efficient apprenticeship-learning algorithms in a number of domains, including two types of relational MDPs. We instantiate our approach in a software agent and a robot agent that learn effectively from a human teacher.
T. Walsh, K. Subramanian, M. Littman, C. Diuk

Toward Off-Policy Learning Control with Function Approximation (377)
We present the first temporal-difference learning algorithm for off-policy control with unrestricted linear function approximation whose per-time-step complexity is linear in the number of features. Our algorithm, text it {Greedy-GQ}, is an extension of recent work on gradient temporal-difference learning, which has hitherto been restricted to a prediction (policy evaluation) setting, to a control setting in which the target policy is greedy with respect to a linear approximation to the optimal action-value function. A limitation of our control setting is that we require the behavior policy to be stationary. We call this setting text it {latent learning} because the optimal policy, though learned, is not manifest in behavior. Popular off-policy algorithms such as Q-learning are known to be unstable in this setting when used with linear function approximation.
H. Maei, C. Szepesvari, S. Bhatnagar, R. Sutton

Efficient Reinforcement Learning with Multiple Reward Functions for Randomized Controlled Trial Analysis (385)
We introduce new, efficient algorithms for value iteration with multiple reward functions and continuous state. We also give an algorithm for finding the set of all non-dominated actions in the continuous state setting. This novel extension is appropriate for environments with continuous or finely discretized states where generalization is required, as is the case for data analysis of randomized controlled trials.
D. Lizotte, M. Bowling, S. Murphy

Internal Rewards Mitigate Agent Boundedness (393)
Reinforcement learning (RL) research typically develops algorithms for helping an RL agent best achieve its goals ---however they came to be defined--- while ignoring the relationship of those goals to the goals of the agent designer. We extend agent design to include the meta-optimization problem of selecting internal agent goals (rewards)which optimize the designer's goals. Our claim is that well-designed internal rewards can help improve the performance of RL agents which are computationally bounded in some way (as practical agents are). We present a formal framework for understanding both bounded agents and the meta-optimization problem, and we empirically demonstrate several instances of common agent bounds being mitigated by general internal reward functions.
J. Sorg, S. Singh, R. Lewis

3c: Deep Learning 2 (Rimon, Tue 15.40)

Restricted Boltzmann Machines are Hard to Approximately Evaluate or Simulate (401)
Restricted Boltzmann Machines (RBMs) are a type of probability modelover the Boolean cube $\{-1,1\}^n$ that have recently received muchattention. We establish the intractability of two basiccomputational tasks involving RBMs, even if only a coarseapproximation to the correct output is required.We first show that assuming P != NP, for any fixed positive constant K (which may be arbitrarily large)there is no polynomial-time algorithm for the followingproblem: given an n-bit input string x and the parameters of aRBM M, output an estimate of the probability assigned to x byM that is accurate to within a multiplicative factor of e^{Kn}. This hardness result holds even if the parametersof M are constrained to be at most $\psi(n)$ for any function $\psi(n) = \omega(n)$, and if the number of hidden nodes ofM is at most n. We then show that assuming RP != NP, there is no polynomial-time randomized algorithm for the following problem: given the parameters of an RBM M, generate arandom example from a probability distribution whose total variationdistance from the distribution defined by M is at most 1/12.
P. Long, R. Servedio

Deep Supervised T-Distributed Embedding (409)
Deep learning has been successfully applied to learn non-linear feature mappings and to perform dimensionality reduction. In this paper, we present supervised embedding techniques that use a deep neural network to collapse classes. The network is pre-trained using a stack of Restricted Boltzmann Machines (RBMs), and finetuned using approaches that try to collapse classes. The finetuning is inspired by ideas from Neighborhood Components Analysis (NCA), but it uses a Student t-distribution to model the probabilities of pairwise data points belonging to the same class in the embedding. We investigate two types of objective functions: deep t-distributed MCML (dt-MCML) and deep t-distributed NCA (dt-NCA). Our experiments on two handwritten digit datasets reveal the strong performance of dt-MCML in supervised parametric data visualization, whereas dt-NCA outperforms alternative techniques when embeddings with more than two or three dimensions are constructed, e.g., to obtain good classification performances. Overall, our results demonstrate the advantage of using a deep architecture and a heavy-tailed t-distribution for measuring pairwise similarities in supervised embedding.
R. Min, L. van der Maaten, Z. Yuan, A. Bonner, Z. Zhang

Rectified Linear Units Improve Restricted Boltzmann Machines (417)
Restricted Boltzmann machines were developed using binary stochastic hidden units. These can be generalized by replacing each binary unit by an infinite number of copies that all have the same weights but have progressively more negative biases. The learning and inference rules for these "Stepped Sigmoid Units" are unchanged. They can be approximated efficiently by noisy, rectified linear units. Compared with binary units, these units learn features that are better for object recognition on the NORB dataset and face verification on the Labeled Faces in the Wild dataset. Unlike binary units, rectified linear units preserve information about relative intensities as information travels through multiple layers of feature detectors.
V. Nair, G. Hinton

Learning Deep Boltzmann Machines using Adaptive MCMC (425)
When modeling high-dimensional richly structured data, it is often the case that the distribution defined by the Deep Boltzmann Machine (DBM) has a rough energy landscape with many local minima that are separated by high energy barriers. The commonly used Gibbs sampler tends to get trapped in one local mode, which often results in unstable learning dynamics and leads to poor parameter estimates. In this paper, we concentrate on learning DBM's using adaptive MCMC algorithms. We first show a close connection between Fast PCD and adaptive MCMC. We then develop a Coupled Adaptive Simulated Tempering algorithm that can be used to better explore a highly multimodal energy landscape. Finally, we demonstrate that the proposed algorithm considerably improves parameter estimates, particularly when learning large-scale DBM's.
R. Salakhutdinov

3d: Structured Output Learning (Hadas, Tue 15.40)

Structured Output Learning with Indirect Supervision (433)
We present a novel approach for structure prediction that addresses the difficulty of obtaining labeled structures for training. We observe that structured output problems often have a companion learning problem of determining whether a given input possesses a good structure. For example, the companion problem for the part-of-speech (POS) tagging task asks whether a given sequence of words has a corresponding sequence of POS tags that is ``legitimate''. While obtaining direct supervision for structures is difficult and expensive, it is often very easy to obtain indirect supervision from the companion binary decision problem.In this paper, we develop a large margin framework that jointly learns from both direct and indirect forms of supervision. Our experiments exhibit the significant contribution of the easy-to-get indirect binary supervision on three important NLP structure learning problems.
M.-W. Chang, V. Srikumar, D. Goldwasser, Dan Roth

Learning Tree Conditional Random Fields (441)
We examine maximum spanning tree-based methods for learning the structure of tree Conditional Random Fields (CRFs) P(Y|X). We use edge weights which take advantage of local inputs X and thus scale to large problems. For a general class of edge weights, we give a negative learn ability result. However, we demonstrate that two members of the class--local Conditional Mutual Information and Decomposable

Conditional Influence--have reasonable theoretical bases and perform very well in practice. On synthetic data and a large-scale fMRI application, our methods outperform existing techniques.
J. Bradley, C. Guestrin

Learning efficiently with approximate inference via dual losses (449)
Many structured prediction tasks involve complex models where inference is computationally intractable, but where it can be well approximated using a linear programming relaxation. Previous approaches for learning for structured prediction (e.g., cutting-plane, subgradient, perceptron) repeatedly make predictions for some of the data points. These approaches are computationally demanding because each prediction involves solving a linear program to optimality.We present a scalable algorithm for learning for structured prediction.The main idea is to instead solve the dual of the structured prediction loss. We formulate the learning task as a convex minimization over both the weights and the dual variables corresponding to each data point. As a result, we can begin to optimize the weights even before completely solving any of the individual prediction problems. We show how the dual variables can be efficiently optimized using coordinate descent. Our algorithm is competitive with state-of-the-art methods such as stochastic subgradient and cutting-plane.
O. Meshi, D. Sontag, T. Jaakkola, A. Globerson

3e: Dimensionality Reduction 1 (Arava, Tue 15.40)

Estimation of (near) low-rank matrices with noise and high-dimensional scaling (457)
We study an instance of high-dimensional statistical inference inwhich the goal is to use N noisy observations to estimate a matrix$\Theta^* \in R^{k \times p}$ that is assumed to be either exactlylow rank, or "near" low-rank, meaning that it can bewell-approximated by a matrix with low rank. We consider an M-estimator based on regularization by the trace or nuclear normover matrices, and analyze its performance under high-dimensionalscaling. We provide non-asymptotic bounds on the Frobenius norm errorthat hold for a general class of noisy observation models, and applyto both exactly low-rank and approximately low-rank matrices. We thenillustrate their consequences for a number of specific learningmodels, including low-rank multivariate or multi-task regression,system identification in vector autoregressive processes, and recoveryof low-rank matrices from random projections. Simulations showexcellent agreement with the high-dimensional scaling of the errorpredicted by our theory.
S. Negahban, M. Wainwright

A DC Programming Approach for Sparse Eigenvalue Problem (465)
We investigate the sparse eigenvalue problem which arises in various fields such as machine learning and statistics. Unlike standard approaches relying on approximation of the l_{0}-norm, we work with an equivalent reformulation of the problem at hand as a DC program. Our starting point is the eigenvalue problem to which a constraint for sparsity requirement is added. The obtained problem is first formulated as a mixed integer program, and exact penalty techniques are used to equivalently transform the resulting problem into a DC program, whose solution is assumed by a customized DCA. Computational results for sparse principal component analysis are reported, which show the usefulness of our approach that compares favorably with some related standard methods using approximation of the l_{0}-norm.
Mamadou Thiao, Tao Pham Dinh, Hoai An Le Thi

A Fast Augmented Lagrangian Algorithm for Learning Low-Rank Matrices (473)
We propose a general and efficient algorithm for learning low-rank matrices. The proposed algorithm converges super-linearly and can keep the matrix to be learned in a compact factorized representation without the need of specifying the rank beforehand. Moreover, we show that the framework can be easily generalized to the problem of learning multiple matrices and general spectral regularization. Empirically we show that we can recover a 10,000x10,000 matrix from 1.2 million observations in about 5 minutes. Furthermore, we show that in a brain-computer interface problem, the proposed method can speed-up the optimization by two orders of magnitude against the conventional projected gradient method and produces more reliable solutions.
R. Tomioka, T. Suzuki, M. Sugiyama, H. Kashima

On Sparse Nonparametric Conditional Covariance Selection (481)

We develop a penalized kernel smoothing method for the problem of selecting non-zero elements of the conditional precision matrix, known as conditional covariance selection. This problem has a key role in many modern applications such as finance and computational biology. However, it has not been properly addressed. Our estimator is derived under minimal assumptions on the underlying probability distribution and works well in the high-dimensional setting. The efficiency of the algorithm is demonstrated on both simulation studies and the analysis of the stock market.
M. Kolar, A. Parikh, E. Xing

4a: Graph Clustering (Alon, Wed 10.30)

Finding Planted Partitions in Nearly Linear Time using Arrested Spectral Clustering (489)
We describe an algorithm for clustering using a similarity graph. Thealgorithm (a) runs in $O(n log^3 n + m \log n)$ time on graphs withn vertices and m edges, and (b) with high probability, finds all``large enough'' clusters in a random graph generated according to theplanted partition model. We provide lower bounds that imply that our``large enough'' constraint cannot be improved much, even using acomputationally unbounded algorithm. We describe some experimentsrunning the algorithm and a few related algorithms on random graphswith partitions generated using a Chinese Restaurant Processes, andsome results of applying the algorithm to cluster DBLP titles.
N. Bshouty, P. Long

Total Variation and Cheeger Cuts (497)
In this work, inspired by (Buhler and Hein, 2009), (Strang, 1983), and (Zhang et al., 2009),we give a continuous relaxation of the Cheeger cut problem on a weighted graph. We show that the relaxation is actually equivalent to the original problem. We then describe an algorithm for finding good cuts suggested by the similarities of the energy of the relaxed problem and various well studied energies in image processing. Finally we provide experimental validation of the proposed algorithm, demonstrating its efficiency in finding high quality cuts.
A. Szlam, X. Bresson

Power Iteration Clustering (505)
We present a simple and scalable graph clustering method called power iteration clustering (PIC). PIC finds a very low-dimensional embedding of a dataset using truncated power iteration on a normalized pair-wise similarity matrix of the data. This embedding turns out to be an effective cluster indicator, consistently outperforming widely used spectral methods such as NCut on real datasets. PIC is very fast on large datasets, running over 1,000 times faster than an NCut implementation based on the state-of-the-art IRAM eigenvector computation technique.
F. Lin, W. Cohen

Robust Graph Mode Seeking by Graph Shift (513)
In this paper, we study how to robustly computethe modes of a graph, namely the densesubgraphs, which characterize the underlyingcompact patterns and are thus useful formany applications. We first define the modesbased on graph density function, then proposethe graph shift algorithm, which startsfrom each vertex and iteratively shifts towardsthe nearest mode of the graph alonga certain trajectory. Both theoretic analysisand experiments show that graph shift algorithmis very efficient and robust, especiallywhen there exist large amount of noises andoutliers.
H. Liu, S. Yan

4b: Reinforcement Learning 4 (Tamar, Wed 10.30)

Analysis of a Classification-based Policy Iteration Algorithm (521)
We present a classification-based policy iteration algorithm, called Direct Policy Iteration, and provide its finite-sample analysis. Our results state a performance bound in terms of the number of policy improvement steps, the number of rollouts used in each iteration, the capacity of the considered policy space, and a new capacity measure which indicates how well the policy space can approximate policies that are greedy w.r.t. any of its members. The analysis reveals a tradeoff between the estimation and

approximation errors in this classification-based policy iteration setting. We also study the consistency of the method when there exists a sequence of policy spaces with increasing capacity.
A. Lazaric, M. Ghavamzadeh, R. Munos

Nonparametric Return Distribution Approximation for Reinforcement Learning (529)
Standard Reinforcement Learning (RL) aims to optimize decision-making rules in terms of the expected return. However, especially for risk-management purposes, other criteria such as the expected shortfall are sometimes preferred. Here, we describe a method of approximating the distribution of returns, which allows us to derive various kinds of information about the returns. We first show that the Bellman equation, which is a recursive formula for the expected return, can be extended to the cumulative return distribution. Then we derive a nonparametric return distribution estimator with particle smoothing based on this extended Bellman equation. A key aspect of the proposed algorithm is to represent the recursion relation in the extended Bellman equation by a simple replacement procedure of particles associated with a state by using those of the successor state. We show that our algorithm leads to a risk-sensitive RL paradigm. The usefulness of the proposed approach is demonstrated through numerical experiments.
T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, T. Tanaka

Inverse Optimal Control with Linearly Solvable MDPs (537)
We present new algorithms for inverse optimal control (or inverse reinforcement learning, IRL) within the framework of linearly-solvable MDPs (LMDPs). Unlike most prior IRL algorithms which recover only the control policy of the expert, we recover the policy, the value function and the cost function. This is possible because here the cost and value functions are uniquely defined given the policy. Despite these special properties, we can handle a wide variety of problems such as the grid worlds popular in RL and most of the nonlinear problems arising in robotics and control engineering. Direct comparisons to prior IRL algorithms show that our new algorithms provide more information and are orders of magnitude faster. Indeed our fastest algorithm is the first inverse algorithm which does not require solving the forward problem; instead it performs unconstrained optimization of a convex and easy-to-compute log-likelihood. Our work also sheds light on the recent Maximum Entropy (MaxEntIRL) algorithm, which was defined in terms of density estimation and the corresponding forward problem was left unspecified. We show that MaxEntIRL is inverting an LMDP, using the less efficient of the algorithms derived here. Unlike all prior IRL algorithms which assume pre-existing features, we study feature adaptation and show that such adaptation is essential in continuous state spaces.
K. Dvijotham, E. Todorov

Feature Selection Using Regularization in Approximate Linear Programs for Markov Decision Processes (545)
Approximate dynamic programming has been used successfully in a large variety of domains, but it relies on a small set of provided approximation features to calculate solutions reliably. Large and rich sets of features can cause the existing algorithms to overfit because of the limited number of samples. We address this shortcoming using L_1 regularization in approximate linear programming. Because the proposed method can automatically select the appropriate richness of features, its performance does not degrade with an increasing number of features. These results rely on new and stronger sampling bounds for regularized approximate linear programs. We also propose a computationally efficient homotopy method. The empirical evaluation of the approach shows that the proposed method performs well on simple MDPs and standard benchmark problems.
M. Petrik, G. Taylor, R. Parr, S. Zilberstein

4c: Risk estimation and Cost-sensitive Learning (Rimon, Wed 10.30)

Risk minimization, probability elicitation, and cost-sensitive SVMs (553)
A new procedure for learning cost-sensitive SVM classifiers is proposed. The SVM hinge loss is extended to the cost sensitive setting, and the cost-sensitive SVM is derived as the minimizer of the associated risk. The extension of the hinge loss draws on recent connections between risk minimization and probability elicitation. These connections are generalized to cost-sensitive classification, in a manner that guarantees

consistency with the cost-sensitive Bayes risk, and associated Bayes decision rule. This ensures that optimal decision rules, under the new hinge loss, implement the Bayes-optimal cost-sensitive classification boundary. Minimization of the new hinge loss is shown to be a generalization of the classic SVM optimization problem, and can be solved by identical procedures. The resulting algorithm avoids the shortcomings of previous approaches to cost-sensitive SVM design, and has superior experimental performance.
Hamed Masnadi-Shirazi, n. Vasconcelos

One-sided Support Vector Regression for Multiclass Cost-sensitive Classification (561)
We propose a novel approach that reduces cost-sensitive classification to one-sided regression. The approach stores the cost information in the regression labels and encodes the minimum-cost prediction with the one-sided loss. The simple approach is accompanied by a solid theoretical guarantee of error transformation, and can be used to cast any one-sided regression method as a cost-sensitive classification algorithm. To validate the proposed reduction approach, we design a new cost-sensitive classification algorithm by coupling the approach with a variant of the support vector machine (SVM) for one-sided regression. The proposed algorithm can be viewed as a theoretically justified extension of the popular one-versus-all SVM. Experimental results demonstrate that the algorithm is not only superior to traditional one-versus-all SVM for cost-sensitive classification, but also better than many existing SVM-based cost-sensitive classification algorithms.
H.-H. Tu, H.-T. Lin

Active Risk Estimation (569)
We address the problem of evaluating the risk of a given model accurately at minimal labeling costs. This problem occurs in situations in which risk estimates cannot be obtained from held-out training data, because the training data are unavailable or do not reflect the desired test distribution. We study active risk estimation processes in which instances are actively selected by a sampling process from a pool of unlabeled test instances and their labels are queried. We derive the sampling distribution that minimizes the estimation error of the active risk estimator when used to select instances from the pool. An analysis of the distribution that governs the estimator leads to confidence intervals. We empirically study conditions under which the active risk estimate is more accurate than a standard risk estimate that draws equally many instances from the test distribution.
C. Sawade, N. Landwehr, S. Bickel, T. Scheffer

Unsupervised Risk Stratification in Clinical Datasets: Identifying Patients at Risk of Rare Outcomes (577)
Most existing algorithms for clinical risk stratification rely on labeled training data. Collecting this data is challenging for clinical conditions where only a small percentage of patients experience adverse outcomes. We propose an unsupervised anomaly detection approach to risk stratify patients without the need of positively and negatively labeled training examples. High risk patients are identified without any expert knowledge using a minimum enclosing ball to find cases that lie in sparse regions of the feature space. When evaluated on data from patients admitted with acute coronary syndrome and patients undergoing inpatient surgical procedures, our approach was able to successfully identify individuals at increased risk of adverse endpoints in both populations. In some cases, unsupervised anomaly detection outperformed other machine learning methods that used additional knowledge in the form of labeled examples.
Z. Syed, I. Rubinfeld

4d: Kernels (Hadas, Wed 10.30)

Two-Stage Learning Kernel Algorithms (585)
This paper examines two-stage techniques for learning kernels based on a notion of alignment. It presents a number of novel theoretical, algorithmic, and empirical results for alignment-based techniques. Our results build on previous work by Cristianini et al. (2001), but we adopt a different definition of kernel alignment and significantly extend that work in several directions: we give a novel and simple concentration bound for alignment between kernel matrices; show the existence of good predictors for kernels with high alignment, both for classification and for regression; give algorithms for learning a maximum alignment kernel by showing that the problem can be reduced to a simple QP; and report the results of extensive

experiments with this alignment-based method in classification and regression tasks, which show an improvement both over the uniform combination of kernels and over other state-of-the-art learning kernel methods.
C. Cortes, M. Mohri, A. Rostamizadeh

Fast Neighborhood Subgraph Pairwise Distance Kernel (593)
We introduce a novel graph kernel called the Neighborhood Subgraph Pairwise Distance Kernel. The kernel decomposes a graph into all pairs of neighborhood subgraphs of small radius at increasing distances. We show that using a fast graph invariant we obtain significant speed-ups in the Gram matrix computation. Finally, we test the novel kernel on a wide range of chemoinformatics tasks, from antiviral to anti carcinogenic to toxicological activity prediction, and observe competitive performance when compared against several recent graph kernel methods.
F. Costa, K. De Grave

On learning with kernels for unordered pairs (601)
We propose and analyze two strategies to learn over unordered pairs with kernels, and provide a common theoretical framework to compare them. The strategies are related to methods that were recently investigated to predict edges in biological networks. We show that both strategies differ in their loss function and in the kernels they use. We deduce in particular a smooth interpolation between the two approaches, as well as new ways to learn over unordered pairs. The different approaches are tested on the inference of missing edges in two biological networks.
M. Hue, J.-P. Vert

Generalization Bounds for Learning Kernels (609)
This paper presents several novel generalization bounds for the problem of learning kernels based on a combinatorial analysis of the Rademacher complexity of the corresponding hypothesis sets. Our bound for learning kernels with a convex combination of p base kernels using L_1 regularization admits only a $\sqrt{\log p}$ dependency on the number of kernels, which is *tight* and considerably more favorable than the previous best bound given for the same problem. We also give a novel bound for learning with a non-negative combination of p base kernels with an L_2 regularization whose dependency on p is also*tight* and only in $p^{1/4}$. We present similar results for L_q regularization with other values of q, and outline the relevance of our proof techniques to the analysis of the complexity of the class of linear functions. Experiments with a large number of kernels further validate the behavior of the generalization error as a function of p predicted by our bounds.
C. Cortes, M. Mohri, A. Rostamizadeh

4e: Dimensionality Reduction 2 (Arava, Wed 10.30)

Local Minima Embedding (617)
Dimensionality reduction is a commonly used step in many algorithms for visualization, classification, clustering and modeling. Most dimensionality reduction algorithms find a low dimensional embedding that preserves the structure of high-dimensional data points. This paper proposes Local Minima Embedding (LME), a technique to find a low-dimensional embedding that preserves the local minima structure of a given objective function. LME provides an embedding that is useful for visualizing and understanding the relation between the original variables that create local minima. Additionally, the embedding can potentially be used to sample the original function to discover new local minima. The function in the embedded space takes an analytic form and hence the gradients can be computed analytically. We illustrate the benefits of LME in both synthetic data and real problems in the context of image alignment. To the best of our knowledge this is the first paper that addresses the problem of finding an embedding that preserves local minima properties of an objective function.
M. Kim, F. De la Torre

Projection Penalties: Dimension Reduction without Loss (625)
Dimension reduction is popular for learning predictive models in high-dimensional spaces. It can highlight the relevant part of the feature space and avoid the curse of dimensionality. However, it can also be harmful because any reduction loses information. In this paper, we propose the projection penalty

framework to make use of dimension reduction without losing valuable information .Reducing the feature space before learning predictive models can be viewed as restricting the model search to some parameter subspace. The idea of projection penalties is that instead of restricting the search to a parameter subspace, we can search in the full space but penalize the projection distance to this subspace. Dimension reduction is used to guide the search, rather than to restrict it.We propose projection penalties for linear dimension reduction, and then generalize to kernel-based reduction and other nonlinear methods. We test projection penalties with various dimension reduction techniques in different prediction tasks, including principal component regression and partial least squares in regression tasks, kernel dimension reduction in face recognition, and latent topic modeling in text classification. Experimental results show that projection penalties are a more effective and reliable way to make use of dimension reduction techniques.
Y. Zhang, J. Schneider

The Elastic Embedding Algorithm for Dimensionality Reduction (633)
We propose a new dimensionality reduction method, the elastic embedding (EE), that optimises an intuitive, nonlinear objective function of the low-dimensional coordinates of the data. The method reveals a fundamental relation betwen a spectral method, Laplacian eigenmaps, and a nonlinear method, stochastic neighbour embedding; and shows that EE can be seen as learning both the coordinates and the affinities between data points. We give a homotopy method to train EE, characterise the critical value of the homotopy parameter, and study the method's behaviour. For a fixed homotopy parameter, we give a globally convergent iterative algorithm that is very effective and requires no user parameters. Finally, we give an extension to out-of-sample points. In standard datasets, EE obtains results as good or better than those of SNE, but more efficiently and robustly.
Miguel Carreira-Perpinan

From Transformation-Based Dimensionality Reduction to Feature Selection (641)
Many learning applications are characterized by high dimensions. Usually not all of these dimensions are relevant and some are redundant. There are two main approaches to reduce dimensionality: feature selection and feature transformation. When one wishes to keep the original meaning of the features, feature selection is desired. Feature selection and transformation are typically presented separately. In this paper, we introduce a general approach for converting transformation-based methods to feature selection methods through L_1/L_∞ regularization. Instead of solving feature selection as a discrete optimization, we relax and formulate the problem as a continuous optimization problem. An additional advantage of our formulation is that our optimization criterion optimizes for feature relevance and redundancy removal automatically. Here, we illustrate how our approach can be utilized to convert linear discriminant analysis (LDA) and the dimensionality reduction version of the Hilbert-Schmidt Independence Criterion (HSIC) to two new feature selection algorithms. Experiments show that our new feature selection methods out-perform related state-of-the-art feature selection approaches.
M. Masaeli, G. Fung, J. Dy

5a: Invited Applications 1 (Alon, Wed 15.40)

Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine (649)
We describe a new Bayesian click-through rate (CTR) prediction algorithm used for Sponsored Search in Microsoft's Bing search engine. The algorithm is based on a probit regression model that maps discrete or real-valued input features to probabilities. It maintains Gaussian beliefs over weights of the model and performs Gaussian online updates derived from approximate message passing. Scalability of the algorithm is ensured through a principled weight pruning procedure and an approximate parallel implementation. We discuss the challenges arising from evaluating and tuning the predictor as part of the complex system of sponsored search where the predictions made by the algorithm decide about future training sample composition. Finally, we show experimental results from the production system and compare to a calibrated Naive Bayes algorithm.
T. Graepel, J. Quinonero Candela, T. Borchert, R. Herbrich

Detecting Large-Scale System Problems by Mining Console Logs (657)
Surprisingly, console logs rarely help operators detect problems in large-scale datacenter services, for they

often consist of the voluminous intermixing of messages from many software components written by independent developers. We propose a general methodology to mine this rich source of information to automatically detect system runtime problems. We use a combination of program analysis and information retrieval techniques to transform free-text console logs into numerical features, which captures sequences of events in the system. We then analyze these features using machine learning to detect operational problems. We also show how to distill the results of our analysis to an operator-friendly one-page decision tree showing the critical messages associated with the detected problems. In addition, we extend our methods to online problem detection where the sequences of events are continuously generated as data streams.
W. Xu, L. Huang, A. Fox, D. Patterson, M. I. Jordan

The Role of Machine Learning in Business Optimization (665)
In a trend that reflects the increasing demand for intelligent applications driven by business data, IBM today is building out a significant number of applications that leverage machine learning technologies to optimize business process decisions. This talk highlights this trend; and describes the many different ways in which leading edge machine learning concepts are being utilized in business applications developed by IBM for its internal use and for clients.
C. Apte

5b: Clustering 1 (Tamar, Wed 15.40)

The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data (667)
We present a probabilistic model for clustering of objects represented via pair wise dissimilarities. We propose that even if an underlying vectorial representation exists, it is better to work directly with the dissimilarity matrix hence avoiding unnecessary bias and variance caused by embeddings.By using a Dirichlet process prior we are not obliged to fix the number of clusters in advance. Furthermore, our clustering model is permutation-, scale- and translation-invariant, and it is called the Translation-invariant Wishart Dirichlet(TIWD) process. A highly efficient MCMC sampling algorithm is presented. Experiments show that the TIWD process exhibits several advantages over competing approaches.
J. Vogt, S. Prabhakaran, T. Fuchs, V. Roth

Clustering processes (675)
The problem of clustering is considered, for the case when each data point is a sample generated by a stationary ergodic process. We propose a very natural asymptotic notion of consistency,and show that simple consistent algorithms exist, under most general non-parametric assumptions. The notion of consistency is as follows: two samples should be put into the same cluster if and only if they we regenerated by the same distribution. With this notion of consistency, clustering generalizes such classical statistical problems as homogeneity testing and process classification. We show that, for the case of a known number of clusters, consistency can be achieved under the only assumption that the joint distribution of the data is stationary ergodic (no parametric or Markovian assumptions, no assumptions of independence, neither between nor within the samples). If the number of clusters is unknown, consistency can be achieved under appropriate assumptions on the mixing rates of the processes (again, no parametric or independence assumptions). In both cases we give examples of simple (at most quadratic in each argument) algorithms which are consistent.
D. Ryabko

Variable Selection in Model-Based Clustering: To Do or To Facilitate (683)
Variable selection for cluster analysis is a difficult problem. The difficulty originates not only from the lack of class information but also the fact that high-dimensional data are often multifaceted and can be meaningfully clustered in multiple ways. In such a case the effort to find one subset of attributes that presumably gives the ``best'' clustering may be misguided. It makes more sense to facilitate variable selection by domain experts, that is, to systematically identify various facets of a data set (each being based on a subset of attributes), cluster the data along each one, and present the results to the domain experts for appraisal and selection. In this paper, we propose a generalization of the Gaussian mixture model, show its ability to cluster data along multiple facets, and demonstrate it is often more reasonable to facilitate variable selection than to perform it.

L. Poon, N. Zhang, T. Chen, Y. Wang

A Nonparametric Information Theoretic Clustering Algorithm (691)
In this paper we propose a novel clustering algorithm based on maximizing the mutual information between data points and clusters. Unlike previous methods, we neitherassume the data are given in terms of distributions nor impose any parametric model on the within-cluster distribution. Instead, we utilize a non-parametric estimation of the average cluster entropies and search for a clustering that maximizes the estimated mutual information between data points and clusters. The improved performance of the proposed algorithm is demonstrated on several standard datasets.
L. Faivishevsky, J. Goldberger

5c: Causal Inference (Rimon, Wed 15.40)

Learning Temporal Graphs for Relational Time-Series Analysis (699)
Learning temporal causal graph structures from multivariate time-series data reveals important dependency relationships between current observations and histories, and provides a better understanding of complex systems. In this paper, we examine learning tasks where one is presented with multiple multivariate time-series, as well as a relational graph among the different time-series. We propose an L1 regularized hidden Markov random field regression framework to leverage the information provided by the relational graph and jointly infer more accurate temporal causal structures for all time-series. We test the proposed model on climate modeling and cross-species micro array data analysis applications.
Y. Liu, Alexandru Niculescu-Mizil, A. Lozano, Y. Lu

Causal filter selection in microarray data (707)
The importance of bringing causality intoplay when designing feature selection methods is more and more acknowledged in the machine learning community. This paper proposes a filter approach based on information theory which aimsto prioritise direct causal relationships in feature selection problems where the ratio between the numberof features and the number of samples is high. This approach is based on thenotion of interaction which is shown to be informative about the relevance of an input subset as well asits causal relationship with the target.The resulting filter, called mIMR (min-Interaction Max-Relevance), is compared with state-of-the-art approaches. Classification results on 25 real microarray datasets show thatthe incorporation of causal aspects in the feature assessment is beneficial both for the resulting accuracy and stability. A toy example of causal discovery shows the effectivenessof the filter for identifying direct causal relationships.
G. Bontempi, P. Meyer

Telling cause from effect based on high-dimensional observations (715)
We describe a method for inferring linear causal relations among multi-dimensional variables. The idea is to use an asymmetry between the distributions of cause and effect that occurs if the covariance matrix of the cause and the structure matrix mapping the cause to the effect are independently chosen.The method applies to both stochastic and deterministic causal relations, provided that the dimensionality is sufficiently high (in some experiments, 5 was enough). It is applicable to Gaussian as well as non-Gaussian data.
D. Janzing, P. Hoyer, B. Scholkopf

Modeling Interaction via the Principle of Maximum Causal Entropy (723)
The principle of maximum entropy provides a powerful framework for statistical models of joint, conditional, and marginal distributions. However, there are many important distributions with elements of interaction and feedback where its applicability has not been established. This work presents the principle of maximum causal entropy--an approach based on causally conditioned probabilities that can appropriately model the availability and influence of sequentially revealed side information. Using this principle, we derive models for sequential data with revealed information, interaction, and feedback, and demonstrate their applicability for statistically framing inverse optimal control and decision prediction tasks.
B. Ziebart, D. Bagnell, A. Dey

5d: Large Margin Methods (Hadas, Wed 15.40)

The Margin Perceptron with Unlearning (731)
We introduce into the classical Perceptron algorithm with margin a mechanism of unlearning which in the course of the regular update allows for a reduction of possible contributions from ``very well classified'' patterns to the weight vector. The resulting incremental classification algorithm, called Margin Perceptron with Unlearning (MPU), provably converges in a finite number of updates to any desirable chosen before running approximation of either the maximal margin or the optimal 1-norm soft margin solution. Moreover, an experimental comparative evaluation involving representative linear Support Vector Machines reveals that the MPU algorithm is very competitive.
C. Panagiotakopoulos, P. Tsampouka

Multi-Class Pegasos on a Budget (739)
When equipped with kernel functions, online learning algorithms are susceptible to the ``curse of kernelization'' that causes unbounded growth in the model size. To address this issue, we present a family of budgeted online learning algorithms for multi-class classification which have constant space and time complexity per update. Our approach is based on the multi-class version of the popular Pegasos algorithm. It keeps the number of support vectors bounded during learning through budget maintenance. By treating the budget maintenance as a source of the gradient error, we prove that the gap between the budgeted Pegasos and the optimal solution directly depends on the average model degradation due to budget maintenance. To minimize the model degradation, we study greedy multi-class budget maintenance methods based on removal, projection, and merging of support vectors. Empirical results show that the proposed budgeted online algorithms achieve accuracy comparable to non-budget multi-class kernelized Pegasos while being extremely computationally efficient.
Z. Wang, K. Crammer, S. Vucetic

COFFIN : A Computational Framework for Linear SVMs (747)
In a variety of applications, kernel machines such as Support Vector Machines (SVMs) have been used with great success often delivering state-of-the-art results. Using the kernel trick, they work on several domains and even enable heterogeneous data fusion by concatenating feature spaces or multiple kernel learning. Unfortunately, they are not suited for truly large-scale applications since they suffer from the curse of supporting vectors, i.e., the speed of applying SVMs decays linearly with the number of support vectors. In this paper we develop COFFIN --- a new training strategy for linear SVMs that effectively allows the use of on demand computed kernel feature spaces and virtual examples in the primal. With linear training and prediction effort this framework leverages SVM applications to truly large-scale problems: As an example, we train SVMs for human splice site recognition involving 50 million examples and sophisticated string kernels. Additionally, we learn an SVM based gender detector on 5 million examples on low-tech hardware and achieve beyond the state-of-the-art accuracies on both tasks. Source code, data sets and scripts are freely available from http://sonnenburgs.de/soeren/coffin .
S. Sonnenburg, V. Franc

Robust Formulations for Handling Uncertainty in Kernel Matrices (755)
We study the problem of uncertainty in the entries of the Kernel matrix, arising in SVM formulation. Using Chance Constraint Programming and a novel large deviation inequality we derive a formulation which is robust to such noise. The resulting formulation applies when the noise is Gaussian, or has finite support. The formulation in general is non-convex, but in several cases of interest it reduces to a convex program. The problem of uncertainty in kernel matrix is motivated from the real world problem of classifying proteins when the structures are provided with some uncertainty. The formulation derived here naturally incorporates such uncertainty in a principled manner leading to significant improvements over the state of the art.
S. Bhadra, S. Bhattacharya, C. Bhattacharyya, Aharon Ben-Tal

5e: Compact Representations (Arava, Wed 15.40)

Learning Fast Approximations of Sparse Coding (763)
In Sparse Coding (SC), input vectors are reconstructed using a sparse linear combination of basis vectors. SC has become a popular method for extracting features from data. For a given input, SC minimizes a quadratic reconstruction error with an L1 penalty term on the code. The process is often too slow for

applications such as real-time pattern recognition. We proposed two versions of a very fast algorithm that produces approximate estimates of the sparse code that can be used to compute good visual features, or to initialize exact iterative algorithms. The main idea is to train a non-linear, feed-forward predictor with a specific architecture and a fixed depth to produce the best possible approximation of the sparse code. A version of the method, which can be seen as a trainable version of Li and Osher's coordinate descent method, is shown to produce approximate solutions with 10 times less computation than Li and Osher's for the same approximation error. Unlike previous proposals for sparse code predictors, the system allows a kind of approximate explaining away to take place during inference. The resulting predictor is differentiable and can be included into globally-trained recognition systems.
K. Gregor, Y. LeCun

Submodular Dictionary Selection for Sparse Representation (771)
We develop an efficient learning framework to construct signal dictionaries for sparse representation by selecting the dictionary columns from multiple candidate bases. By sparse, we mean that only a few dictionary elements, compared to the ambient signal dimension, can exactly represent or well-approximate the signals of interest. We formulate both the selection of the dictionary columns and the sparse representation of signals as a joint combinatorial optimization problem. The proposed combinatorial objective maximizes variance reduction over the set of training signals by constraining the size of the dictionary as well as the number of dictionary columns that can be used to represent each signal. We show that if the available dictionary column vectors are incoherent, our objective function satisfies approximate submodularity. We exploit this property to develop SDSOMP and SDSMA, two greedy algorithms with approximation guarantees. We also describe how our learning framework enables dictionary selection for structured sparse representations, e.g., where the sparse coefficients occur in restricted patterns. We evaluate our approach on synthetic signals and natural images for representation and in painting problems.
A. Krause, V. Cevher

Proximal Methods for Sparse Hierarchical Dictionary Learning (779)
We propose to combine two approaches for modeling data admitting sparse representations: on the one hand, dictionary learning has proven effective for various signal processing tasks. On the other hand, recent work on structured sparsity provides a natural framework for modeling dependencies between dictionary elements. We thus consider a tree-structured sparse regularization to learn dictionaries embedded in a hierarchy. The involved proximal operator is computable exactly via a primal-dual method, allowing the use of accelerated gradient techniques. Experiments show that for natural image patches, learned dictionary elements organize themselves in such a hierarchical structure, leading to an improved performance for restoration tasks. When applied to text documents, our method learns hierarchies of topics, thus providing a competitive alternative to probabilistic topic models.
R. Jenatton, J. Mairal, G. Obozinski, F. Bach

Sequential Projection Learning for Hashing with Compact Codes (787)
Hashing based Approximate Nearest Neighbor (ANN) search has attracted much attention due to its fast query time and drastically reduced storage. However, most of the hashing methods either use random projections or extract principal directions from the data to derive hash functions. The resulting embedding suffers from poor discrimination when compact codes are used. In this paper, we propose a novel data-dependent projection learning method such that each hash function is designed to correct the errors made by the previous one sequentially. The proposed method easily adapts to both unsupervised and semi-supervised scenarios and shows significant performance gains over the state-of-the-art methods on two large datasets containing up to 1 million points.
J. Wang, S. Kumar, S.-F. Chang

6a: Graphical Models (Alon, Thu 10.30)

Heterogeneous Continuous Dynamic Bayesian Networks with Flexible Structure and Inter-Time Segment Information Sharing (795)
Classical dynamic Bayesian networks (DBNs) are based on the homogeneous Markov assumption and cannot deal with heterogeneity and non-stationarity in temporal processes. Various approaches to relax the homogeneity assumption have recently been proposed. The present paper aims to improve the

shortcomings of three recent versions of heterogeneous DBNs along the following lines:(i) avoiding the need for data discretization,(ii) increasing the flexibility over a time-invariant network structure,(iii) avoiding over-flexibility and over fitting by introducing a regularization scheme based in inter-time segment information sharing. The improved method is evaluated on synthetic data and compared with alternative published methods on gene expression time series from Drosophila melanogaster.
F. Dondelinger, S. Lebre, D. Husmeier

Continuous-Time Belief Propagation (803)
Many temporal processes can be naturally modeled as a stochastic system that evolves continuously over time. The representation language of continuous-time Bayesian networks allows to succinctly describe multi-component continuous-time stochastic processes. A crucial element in applications of such models is inference. Here we introduce a variational approximation scheme, which is a natural extension of Belief Propagation for continuous- time processes. In this scheme, we view messages as inhomogeneous Markov processes over individual components. This leads to a relatively simple procedure that allows to easily incorporate adaptive ordinary differential equation (ODE) solvers to perform individual steps. We provide the theoretical foundations for the approximation, and show how it performs on a range of networks. Our results demonstrate that our method is quite accurate on singly connected networks, and provides close approximations in more complex ones.
T. El-Hay, I. Cohn, N. Friedman, R. Kupferman

Exploiting Data-Independence for Fast Belief-Propagation (811)
Maximum a posteriori (MAP) inference in graphical models requires that we maximize the sum of two terms: a data-dependent term, encoding the conditional likelihood of a certain labeling given an observation, and a data-independent term, encoding some prior on labelings. Often, data-dependent factors contain fewer latent variables than data-independent factors -- for instance, many grid and tree-structured models contain only first-order conditionals despite having pair wise priors. In this paper, we note that MAP-inference in such models can be made substantially faster by appropriately preprocessing their data-independent terms. Our main result is to show that message-passing in any such pair wise model has an expected-case exponent of only 1.5 on the number of states per node, leading to significant improvements over existing quadratic-time solutions.
J. McAuley, T. Caetano

Accelerated dual decomposition for MAP inference (819)
Approximate MAP inference in graphical models is an important and challenging problem for many domains including computer vision, computational biology and natural language understanding. Current state-of-the-art approaches employ convex relaxations of these problems as surrogate objectives, but only provide weak running time guarantees. In this paper, we develop an approximate inference algorithm that is both efficient and has strong theoretical guarantees. Specifically, our algorithm is guaranteed to converge to an ϵ-accurate solution of the convex relaxation in $O\left(\frac{1}{\epsilon}\right)$ time. We demonstrate our approach on synthetic and real-world problems and show that it outperforms current state-of-the-art techniques.
V. Jojic, S. Gould, D. Koller

6b: Clustering 2 (Tamar, Thu 10.30)

Multiple Non-Redundant Spectral Clustering Views (827)
Many clustering algorithms only find one clustering solution. However, data can often be grouped and interpreted in many different ways. This is particularly true in the high-dimensional setting where different subspaces reveal different possible groupings of the data. Instead of committing to one clustering solution, here we introduce a novel method that can provide several non-redundant clustering solutions to the user. Our approach simultaneously learns non-redundant subspaces that provide multiple views and finds a clustering solution in each view. We achieve this by augmenting a spectral clustering objective function to incorporate dimensionality reduction and multiple views and to penalize for redundancy between the views.
D. Niu, J. Dy, M. Jordan

Mining Clustering Dimensions (835)

Although it is common practice to produce only a single clustering of a dataset, in many cases text documents can be clustered along different dimensions. Unfortunately, not only do traditional clustering algorithms fail to produce multiple clusterings of a dataset, the only clustering they produce may not be the one that the user desires. To address this major limitation, we propose a novel clustering algorithm for inducing multiple clusterings along the important dimensions of a dataset. Its ability to reveal the important clustering dimensions of a dataset in an un-supervised manner is particularly appealing for those users who have no idea of how a data can possibly be clustered. We demonstrate its viability on several challenging text classification tasks.
S. Dasgupta, V. Ng

Comparing Clusterings in Space (843)
This paper proposes a new method for comparing clusterings both partitionally and geometrically. Our approach is motivated by the following observation: the vast majority of previous techniques for comparing clusterings are entirely partitional, i.e., they examine assignments of points in set theoretic terms after they have been partitioned. In doing so, these methods ignore the spatial layout of the data, disregarding the fact that this information is responsible for generating the clusterings to begin with. We demonstrate that this leads to a variety of failure modes. Previous comparison techniques often fail to differentiate between significant changes made in data being clustered. We formulate a new measure for comparing clusterings that combines spatial and partitional information into a single measure using optimization theory. Doing so eliminates pathological conditions in previous approaches. It also simultaneously removes common limitations, such as that each clustering must have the same number of clusters or the yare over identical datasets. This approach is stable, easily implemented, and has strong intuitive appeal.
M. Coen, H. Ansari, N. Fillmore

Robust Subspace Segmentation by Low-Rank Representation (851)
We propose low-rank representation(LRR) to segment data drawn from a union of multiple linear (or affine) subspaces. Given a set of data vectors, LRR seeks the lowest-rank representation among all the candidates that represent all vectors as the linear combination of the bases in a dictionary. Unlike the well-known sparse representation (SR), which computes the sparsest representation of each data vector individually, LRR aims at finding the lowest-rank representation of a collection of vectors jointly. LRR better captures the global structure of data, giving a more effective tool for robust subspace segmentation from corrupted data. Both theoretical and experimental results show that LRR is a promising tool for subspace segmentation.
G. Liu, Z. Lin, Y. Yu

6c: Feature and Kernel Selection (Rimon, Thu 10.30)

Simple and Efficient Multiple Kernel Learning By Group Lasso (859)
We consider the problem of how to improve the efficiency of Multiple Kernel Learning (MKL). MKL is often regarded as a convex-concave optimization problem, which is convex on the kernel weights and concave on the SVM dual variables. In literature, the alternating approach is widely observed: (1) the minimization of the kernel weights is solved by complicated techniques, such as Semi-infinite Linear Programming, Gradient Descent, or Level method; (2) the maximization of SVM dual variables can be solved by standard SVM solvers. However, the minimization over kernel weights in these methods is usually dependent on its solving techniques or commercial softwares, which therefore limits the efficiency and applicability. In this paper, we formulate a closed-form solution for the optimization of kernel weights based on the equivalence between group-lasso and MKL. Although this equivalence is not our invention, our derived variant equivalence not only lead to an efficient algorithm for MKL, but also generalize this equivalence to that between Lp-MKL (p >=1 and denoting the Lp-norm of kernel weights) and group lasso, which does not appear in literature. Therefore, our proposed algorithm provides a unified solution for the whole family of Lp-MKL models. Experiments on multiple data sets show the promising performance of the proposed technique compared with other competitive methods.
Z. Xu, R. Jin, H. Yang, I. King, M. Lyu

Online Streaming Feature Selection (867)
We study an interesting and challenging problem, online streaming feature selection, in which the size of the feature set is unknown, and not all features are available for learning while leaving the number of

observations constant. In this problem, the candidate features arrive one at a time, and the learner's task is to select a ``best so far'' set of features from streaming features. Standard feature selection methods cannot perform well in this scenario. Thus, we present a novel framework based on feature relevance. Under this framework, a promising alternative method, Online Streaming Feature Selection (OSFS), is presented to online select strongly relevant and non-redundant features. In addition to OSFS, a faster Fast-OSFS algorithm is proposed to further improve the selection efficiency. Experimental results show that our algorithms achieve more compactness and better accuracy than existing streaming feature selection algorithms on various datasets.
X. Wu, K. Yu, H. Wang, W. Ding

Feature Selection as a One-Player Game (875)
This paper formalizes Feature Selection as a Reinforcement Learning problem, leading to a provably optimal though intractable selection policy. As a second contribution, this paper presents an approximation thereof, based on a one-player game approach and relying on the Monte-Carlo tree search UCT (Upper Confidence Tree) proposed by Kocsis and Szepesvari (2006).More precisely, the presented FUSE (Feature Uct SElection) algorithm extends UCT to deal with i) a finite unknown horizon (the target number of relevant features); ii) a huge branching factor of the search tree (the size of the initial feature set). Additionally, a frugal reward function is proposed as a rough but unbiased estimate of the relevance of a feature subset. A proof of concept of FUSE is shown on benchmark data sets.
R. Gaudel, M. Sebag

Learning Sparse SVM for Feature Selection on Very High Dimensional Datasets (883)
A sparse representation of Support vector Machines (SVMs) with respect to input features is desirable for many applications. In this paper, by introducing a 0-1 control variable to each input feature, l_0-norm Sparse SVM (SSVM) is converted to a mixed integer programming (MIP) problem. Rather than directly solving this MIP, we propose an efficient cutting plane algorithm combining with multiple kernel learning to solve its convex relaxation. A global convergence proof for our method is also presented. Comprehensive experimental results on one synthetic and 10 real world datasets show that our proposed method can obtain better or competitive performance compared with existing SVM-based feature selection methods in term of sparsity and eneralization performance. Moreover, our proposed method can effectively handle large-scale and extremely high dimensional problems.
M. Tan, L. Wang, I. Tsang

6d: Learning Theory (Hadas, Thu 10.30)

Efficient Learning with Partially Observed Attributes (891)
We describe and analyze efficient algorithms for learning a linear predictor from examples when the learner can only view a few attributes of each training example. This is the case, for example, in medical research, where each patient participating in the experiment is only willing to go through a small number of tests. Our analysis bounds the number of additional examples sufficient to compensate for the lack of full information on each training example. We demonstrate the efficiency of our algorithms by showing that when running on digit recognition data, they obtain a high prediction accuracy even when the learner gets to see only four pixels of each image.
N. Cesa-Bianchi, S. Shalev-Shwartz, O. Shamir

On the Interaction between Norm and Dimensionality: Multiple Regimes in Learning (899)
A learning problem might have several measures of complexity (e.g., norm and dimensionality) that affect the generalization error. What is the interaction between these complexities? Dimension-free learning theory bounds and parametric asymptotic analyses each provide a partial picture of the full learning curve. In this paper, we use high-dimensional asymptotics on two classical problems---mean estimation and linear regression---to explore the learning curve more completely. We show that these curves exhibit multiple regimes, where in each regime, the excess risk is controlled by a subset of the problem complexities.
P. Liang, N. Srebro

An Analysis of the Convergence of Graph Laplacians (907)
Existing approaches to analyzing the asymptotics of graph Laplacians typically assume a well-behaved

kernel function with smoothness assumptions. We remove the smoothness assumption and generalize the analysis of graph Laplacians to include previously unstudied graphs including kNN graphs. We also introduce a kernel-free framework to analyze graph constructions with shrinking neighborhoods in general and apply it to analyze locally linear embedding (LLE). We also describe how, for a given limit operator, desirable properties such as a convergent spectrum and sparseness can be achieved by choosing the appropriate graph construction
D. Ting, L. Huang, M. Jordan

6e: Exploration and Feature Construction (Arava, Thu 10.30)

Efficient Selection of Multiple Bandit Arms: Theory and Practice (915)
We consider the general, widely applicable problem of selecting from n real-valued random variables a subset of size m of those with the highest means, based on as few samples as possible. This problem,which we denote Explore-m, is a core aspect in several stochastic optimization algorithms, and applications of simulation and industrial engineering. The theoretical basis for our work is an extension of a previous formulation using multi-armed bandits that is devoted to identifying just the one best of n random variables(Explore-1). In addition to providing PAC bounds for the general case, we tailor our theoretically grounded approach to work efficiently in practice. Empirical comparisons of the resulting sampling algorithm against state-of-the-art subset selection strategies demonstrate significant gains in sample efficiency.
S. Kalyanakrishnan, P. Stone

Model-based reinforcement learning with nearly tight exploration complexity bounds (923)
One might believe that model-based algorithms of reinforcement learning can propagate the obtained experience more quickly, and are able to direct exploration better. As a consequence, fewer exploratory actions should be enough to learn a good policy. Strangely enough, current theoretical results for model-based algorithms do not support this claim: In a finite Markov decision process with N states, the best bounds on the number of exploratory steps necessary are of order $O(N^2 \log N)$, in contrast to the $O(N \log N)$ bound available for the model-free, delayed Q-learning algorithm. In this paper we show that MoRmax, a modified version of the Rmax algorithm needs to make at most $O(N \log N)$ exploratory steps. This matches the lower bound up to logarithmic factors, as well as the upper bound of the state-of-the-art model-free algorithm, while our new bound improves the dependence on other problem parameters.
I. Szita, C. Szepesvari

A theoretical analysis of feature pooling in vision algorithms (931)
Many modern visual recognition algorithms incorporate a step of spatial `pooling', where the outputs of several nearby feature detectors are combined into a local or global `bag of features', in a way that preserves task-related information while removing irrelevant details. Pooling is used to achieve invariance to image transformations, more compact representations, and better robustness to noise and clutter. Several papers have shown that the details of the pooling operation can greatly influence the performance, but studies have so far been purely empirical. In this paper, we show that the reasons underlying the performance of various pooling methods are obscured by several confounding factors, such as the link between the sample cardinality in a spatial pool and the resolution at which low-level features have been extracted. We provide a detailed theoretical analysis of max pooling and average pooling, and give extensive empirical comparisons for object recognition tasks.
Y.-L. Boureau, J. Ponce, Y. LeCun

Improved Local Coordinate Coding using Local Tangents (939)
Local Coordinate Coding (LCC), introduced in (Yu et al., 2009), is a high dimensional nonlinear learning method that explicitly takes advantage of the geometric structure of the data. Its successful use in the winning system of last year's Pascal image classification Challenge (Everingham, 2009) shows that the ability to integrate geometric information is critical for some real world machine learning applications. This paper further develops the idea of integrating geometry in machine learning by extending the original LCC method to include local tangent directions. These new correction terms lead to better approximation of high dimensional nonlinear functions when the underlying data manifold is locally flat. The method significantly reduces the number of anchor points needed in LCC, which not only reduces computational cost, but also improves prediction performance. Experiments are included to demonstrate that this method is more

effective than the original LCC method on some image classification tasks.
K. Yu, T. Zhang

7a: Invited Applications 2 (Alon, Thu 13.30)

Music Plus One and Machine Learning (947)
A system for musical accompaniment is presented in which a computer-driven orchestra follows and learns from a soloist in a concerto-like setting. The system is decomposed into three modules: the first computes a real-time score match using a hidden Markov model; the second generates the output audio by phase-vocoding a preexisting audio recording; the third provides a link between these two, by predicting future timing evolution using a Kalman filter-like model. Several examples are presented showing the system in action in diverse musical settings. Connections with machine learning are highlighted, showing current weaknesses and new possible directions.
C. Raphael

Climbing the Tower of Babel: Unsupervised Multilingual Learning (955)
For centuries, scholars have explored the deep links among human languages. In this paper, we present a class of probabilistic models that use these links as a form of naturally occurring supervision. These models allow us to substantially improve performance for core text processing tasks, such as morphological segmentation, part-of-speech tagging, and syntactic parsing. Besides these traditional NLP tasks, we also present a multilingual model for the computational decipherment of lost languages.
B. Snyder, R. Barzilay

FAB-MAP: Appearance-Based Place Recognition and Mapping using a Learned Visual Vocabulary Model (963)
We present an overview of FAB-MAP, an algorithm for place recognition and mapping developed for infrastructure-free mobile robot navigation in large environments. The system allows a robot to identify when it is revisiting a previously seen location, on the basis of imagery captured by the robot's camera. We outline a complete probabilistic framework for the task, which is applicable even in visually repetitive environments where many locations may appear identical. Our work introduces a number of technical innovations - notably we demonstrate that place recognition performance can be improved by learning an approximation to the joint distribution over visual elements. We also investigate several principled approaches to making the system robust in visually repetitive environments, and define an efficient bail-out strategy for multi-hypothesis testing to improve system speed. Our model has been shown to substantially outperform standard tf-idf ranking on our task of interest. We demonstrate the system performing reliable online appearance mapping and loop closure detection over a 1,000km trajectory, with mean filter update times of 14 ms.
M. Cummins, P. Newman

Discriminative Latent Variable Models for Object Detection (971)
In this talk, I will discuss recent work by colleagues and myself on discriminative latent-variable models for object detection. Object recognition is one of the fundamental challenges of computer vision. We specifically consider the task of localizing and detecting instances of a generic object category, such as people or cars, in cluttered real-word images. Recent benchmark competitions such as the PASCAL Visual Object Challenge suggest our method is the state-of-the-art system for such tasks. This success, combined with publically-available code that runs orders of magnitude faster than comparable approaches, has turned our system into a standard baseline for contemporary research on object recognition (Felzenszwalb et al., 2008; 2009).
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan

7b: Semi-Supervised Learning 1 (Tamar, Thu 13.30)

Large Graph Construction for Scalable Semi-Supervised Learning (972)
In this paper, we address the scalability issue plaguing graph-based semi-supervised learning viaa small number of anchor points which adequately cover the entire point cloud. Critically, these anchor points enable nonparametric regression that predicts the label for each data point as a locally weighted average of

the labels on anchor points. Because conventional graph construction is inefficient in large scale, we propose to construct a tractable large graph by couplinganchor-based label prediction and adjacency matrix design. Contrary to the Nystrom approximation of adjacency matrices which results in indefinite graph Laplacians and in turn leads to potential non-convex optimization over graphs, the proposed graph construction approach based on a unique idea called AnchorGraph provides nonnegative adjacency matrices to guarantee positive semidefinite graph Laplacians. Our approach scales linearly with the data size and in practice usually produces a large sparse graph. Experiments on large datasets demonstrate the significant accuracy improvement and scalability of the proposed approach.
W. Liu, J. He, S.-F. Chang

Multiscale Wavelets on Trees, Graphs and High Dimensional Data: Theory and Applications to Semi Supervised Learning (980)
Harmonic analysis, and in particular the relation between function smoothnessand approximate sparsity of its wavelet coefficients, has played a key role insignal processing and statistical inference for low dimensional data.In contrast, harmonic analysis has thus far had little impact in modern problemsinvolving high dimensional data, or data encoded as graphs or networks.The main contribution of this paper is the development of a harmonic analysisapproach, including both learning algorithms and supporting theory, applicable to these moregeneral settings. Given data (be it high dimensional, graph or network) that is representedby one or more hierarchical trees, we first construct multiscale wavelet-likeorthonormal baseson it.Second, we prove that in analogyto the Euclidean case, function smoothness with respectto a specific metric induced by the tree is equivalent to exponential rate of coefficient decay,that is, to approximate sparsity.These results readily translate to simple practicalalgorithms for various learning tasks.
M. Gavish, B. Nadler, R. Coifman

High-Performance Semi-Supervised Learning using Discriminatively Constrained Generative Models (988)
We develop a semi-supervised learning method that constrains the posterior distribution of latent variables under a generative model to satisfy a rich set of feature expectation constraints estimated with labeled data. This approach encourages the generative model to discover latent structure that is relevant to a prediction task. We estimate parameters with a coordinate ascent algorithm, one step of which involves training a discriminative log-linear model with an embedded generative model. This hybrid model can be used for test time prediction. Unlike other high-performance semi-supervised methods, the proposed method converges to a local maximum of a single objective function, and affords additional flexibility, for example to use different latent and output spaces. We conduct experiments on three sequence labeling tasks, achieving the best reported results on two of them, and showing promising results on CoNLL03 NER.
G. Druck, A. McCallum

Asymptotic Analysis of Generative Semi-Supervised Learning (996)
Semi-supervised learning has emerged as a popular framework for improving modeling accuracy while controlling labeling cost. Based on an extension of stochastic composite likelihood we quantify the asymptotic accuracy of generative semi-supervised learning. In doing so, we complement distribution-free analysis by providing an alternative framework to measure the value associated with different labeling policies and resolve the fundamental question of how much data to label and in what manner. We demonstrate our approach with both simulation studies and real world experiments using naive Bayes for text classification and MRFs and CRFs for structured prediction in NLP.
J. Dillon, K. Balasubramanian, G. Lebanon

7c: Gaussian Processes (Rimon, Thu 13.30)

Sparse Gaussian Process Regression via L_1 Penalization (1004)
To handle massive data, a variety of sparse Gaussian Process (GP) methods have been proposed to reduce the computational cost. Many of them essentially map the large dataset into a small set of basis points. A common approach to learn these basis points is evidence maximization. Nevertheless, maximized evidence may lead to over fitting and cause high computational cost for optimization. In this paper, we propose a novel sparse GP regression approach, GPLasso, that explicitly represents the trade-off between its approximation quality and the model sparsity. GPLasso minimizes a ℓ_1-penalized KL divergence

between the exact and sparse GP posterior processes. Optimizing this convex cost function leads to sparse GP parameters. Furthermore, we use low-rank matrix approximation (for example, incomplete Cholesky factorization) in our optimization procedure to significantly reduce the computational cost. Experimental results demonstrate that, compared with several state-of-the-art sparse GP methods and direct low-rank matrix approximation methods, GPLasso achieves a much improved balance between prediction accuracy and computational cost.
F. Yan, Y. Qi

Surrogating the surrogate: accelerating Gaussian-process-based global optimization with a mixture cross-entropy algorithm (1012)
In global optimization, when the evaluation of the target function is costly, the usual strategy is to learn a surrogate model for the target function and replace the initial optimization by the optimization of the model. Gaussian processes have been widely used since they provide an elegant way to model the fitness and to deal with the exploration-exploitation trade-off in a principled way. Several empirical criteria have been proposed to drive the model optimization, among which is the well-known Expected Improvement criterion. The major computational bottleneck of these algorithms is the exhaustive grid search used to optimize the highly multi modal merit function. In this paper, we propose a competitive ``adaptive grid'' approach, based on a properly derived Cross-Entropy optimization algorithm with mixture proposals. Experiments suggest that 1) we outperform the classical single-Gaussian cross-entropy method when the fitness function is highly multi modal, and 2) we improve on standard exhaustive search in GP-based surrogate optimization.
R. Bardenet, B. Kegl

Gaussian Processes Multiple Instance Learning (1020)
This paper proposes a multiple instance learning (MIL) algorithm for Gaussian processes (GP). The GP-MIL model inherits two crucial benefits from GP: (i) a principle manner of learning kernel parameters, and (ii) a probabilistic interpretation (e.g., variance in prediction) that is informative for better understanding of the MIL prediction problem. The bag labeling protocol of the MIL problem, namely the existence of a positive instance in a bag, can be effectively represented by a sigmoid likelihood model through the max function over GP latent variables. To circumvent the intractability of exact GP inference and learning incurred by the non-continuous max function, we suggest two approximations: first, the soft-max approximation; second, the use of witness indicator variables optimized with a deterministic annealing schedule. The effectiveness of GP-MIL against other state-of-the-art MIL approaches is demonstrated on several benchmark MIL datasets.
M. Kim, F. De la Torre

Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design (1028)
Many applications require optimizing an unknown, noisy function that is expensive to evaluate. We formalize this task as a multi-armed bandit problem, where the payoff function is either sampled from a Gaussian process (GP) or has low RKHS norm. We resolve the important open problem of deriving regret bounds for this setting, which imply novel convergence rates for GP optimization. We analyze GP-UCB, an intuitive upper-confidence based algorithm, and bound its cumulative regret in terms of maximal information gain, establishing a novel connection between GP optimization and experimental design. Moreover, by bounding the latter in terms of operator spectra, we obtain explicit sublinear regret bounds for many commonly used covariance functions. In some important cases, our bounds have surprisingly weak dependence on the dimensionality. In our experiments on real sensor data, GP-UCB compares favorably with other heuristical GP optimization approaches.
N. Srinivas, A. Krause, S. Kakade, M. Seeger

7d: Online Learning (Hadas, Thu 13.30)

Online Prediction with Privacy (1036)
In this paper, we consider online prediction from expert advice in a situation where each expert observes its own loss at each time while the loss cannot be disclosed to others for reasons of privacy or confidentiality preservation. Our secure exponential weighting scheme enables exploitation of such private loss values by making use of cryptographic tools. We proved that the regret bound of the secure

exponential weighting is the same or almost the same with the well-known exponential weighting scheme in the full information model. In addition, we prove theoretically that the secure exponential weighting is privacy-preserving in the sense of secure function evaluation.
J. Sakuma, H. Arai

Implicit Online Learning (1044)
Online learning algorithms have recently risen to prominence due to their strong theoretical guarantees and an increasing number of practical applications for large-scale data analysis problems. In this paper, we analyze a class of online learning algorithms based on fixed potentials and non-linearized losses, which yields algorithms with implicit update rules. We show how to efficiently compute these updates, and we prove regret bounds for the algorithms. We apply our formulation to several special cases where our approach has benefits over existing online learning methods. In particular, we provide improved algorithms and bounds for the online metric learning problem, and show improved robustness for online linear prediction problems. Results over a variety of data sets demonstrate the advantages of our framework.
B. Kulis, P. Bartlett

Online Learning for Group Lasso (1052)
We develop a novel online learning algorithm for the group lasso in order to efficiently find the important explanatory factors in a grouped manner. Different from traditional batch-mode group lasso algorithms, which suffer from the inefficiency and poor scalability, our proposed algorithm performs in an online mode and scales well: at each iteration one can update the weight vector according to a closed-form solution based on the average of previous subgradients. Therefore, the proposed online algorithm can be very efficient and scalable. This is guaranteed by its low worst-case time complexity and memory cost both in the order of ${\cal O}(d)$, where d is the number of dimensions. Moreover, in order to achieve more sparsity in both the group level and the individual feature level, we successively extend our online system to efficiently solve a number of variants of sparse group lasso models. We also show that the online system is applicable to other group lasso models, such as the group lasso with overlap and graph lasso. Finally, we demonstrate the merits of our algorithm by experimenting with both synthetic and real-world datasets.
H. Yang, Z. Xu, I. King, M. Lyu

Random Spanning Trees and the Prediction of Weighted Graphs (1060)
We show that the mistake bound for predicting the nodes of an arbitrary weighted graphics characterized (up to logarithmic factors) by the weighted cut size of a random spanning tree of the graph.The cut size is induced by the unknown adversarial labeling of the graph nodes.In deriving our characterization, we obtain a simple randomized algorithm achieving the optimal mistake bound on any weighted graph.Our algorithm draws a random spanning tree of the original graph and then predicts the nodes of this tree in constant amortized time and linear space. Experiments on real-world diastases that our method compares well to both global (Perceptron) and local (label-propagation) methods, while being much faster.
N. Cesa-Bianchi, C. Gentile, F. Vitale, G. Zappella

7e: Multi-Agent Learning (Arava, Thu 13.30)

Multi-agent Learning Experiments on Repeated Matrix Games (1068)
This paper experimentally evaluates multi-agent learning algorithms playing repeated matrix games to maximize their cumulative return. Previous works assessed that Q-learning surpassed Nash-based multi-agent learning algorithms. Based on all-against-all repeated matrix game tournaments, this paper updates the state of the art of multi-agent learning experiments. In a first stage, it shows that M-Qubed, S and bandit-based algorithms such as UCB are the best algorithms on general-sum games, Exp3 being the best on cooperative games and zero-sum games. In a second stage, our experiments show that two features - forgetting the far past, and using recent history with states - improve the learning algorithms. Finally, the best algorithms are two new algorithms, Q-learning and UCB enhanced with the two features, and M-Qubed.
B. Bouzy, M. Metivier

Classes of Multiagent Q-learning Dynamics with epsilon-greedy Exploration (1076)
Q-learning in single-agent environments is known to converge in the limit given sufficient exploration. The

same algorithm has been applied, with some success, in multiagent environments, where traditional analysis techniques break down. Using established dynamical systems methods, we derive and study an idealization of Q-learning in 2-player 2-action repeated general-sum games. In particular, we address the discontinuous case of epsilon-greedy exploration and use it as a proxy for value-based algorithms to highlight a contrast with existing results in policy search. Analogously to previous results for gradient ascent algorithms, we provide a complete catalog of the convergence behavior of the epsilon-greedy Q-learning algorithm by introducing new subclasses of these games. We identify two subclasses of Prisoner's Dilemma-like games where the application of Q-learning with epsilon-greedy exploration results in higher-than-Nash payoffs for some initial conditions.
M. Wunder, M. Littman, M. Babes

Multiagent Inductive Learning: an Argumentation-based Approach (1084)
Multiagent Inductive Learning is the problem that groups of agents face when they want to perform inductive learning, but the data of interest is distributed among them. This paper focuses on concept learning, and presents A-MAIL, a framework for multiagent induction integrating ideas from inductive learning, case-based reasoning and argumentation. Argumentation is used as a communication framework with which the agents can communicate their inductive inferences to reach shared and agreed-upon concept definitions. We also identify the requirements for learning algorithms to be used in our framework, and propose an algorithm which satisfies them.
Santiago Ontanon, E. Plaza

Convergence, Targeted Optimality, and Safety in Multiagent Learning (1092)
This paper introduces a novel multiagent learning algorithm, Convergence with Model Learning and Safety (or CMLeS in short), which achieves convergence, targeted optimality against memory-bounded adversaries, and safety, in arbitrary repeated games. The most novel aspect of CMLeS is the manner in which it guarantees(in a PAC sense) targeted optimality against memory-bounded adversaries, via efficient exploration and exploitation. CMLeS is fully implemented and we present empirical results demonstrating its effectiveness.
D. Chakraborty, P. Stone

8a: Graphical Models and Bayesian Methods (Alon, Thu 15.40)

Particle Filtered MCMC-MLE with Connections to Contrastive Divergence (1100)
Learning undirected graphical models such as Markov random fields is an important machine learning task with applications in many domains. Since it is usually intractable to learn these models exactly, various approximate learning techniques have been developed, such as contrastive divergence (CD) and Markov chain Monte Carlo maximum likelihood estimation (MCMC-MLE). In this paper, we introduce particle filtered MCMC-MLE, which is a sampling-importance-re sampling version of MCMC-MLE with additional MCMC rejuvenation steps. We also describe a unified view of (1) MCMC-MLE, (2) our particle filtering approach, and (3) a stochastic approximation procedure known as persistent contrastive divergence. We show how these approaches are related to each other and discuss the relative merits of each approach. Empirical results on various undirected models demonstrate that the particle filtering technique we propose in this paper can significantly outperform MCMC-MLE. Furthermore, in certain cases, the proposed technique is faster than persistent CD.
A. Asuncion, Q. Liu, A. Ihler, P. Smyth

Non-Local Contrastive Objectives (1108)
Pseudo-likelihood and contrastive divergence are two well-known examples of contrastive methods. These algorithms trade off the probability of the correct label with the probabilities of other ``nearby'' instantiations. In this paper we explore more general types of contrastive objectives, which trade off the probability of the correct label against an arbitrary set of other instantiations. We prove that a large class of contrastive objectives are consistent with maximum likelihood,even for finite amounts of data. This result generalizes asymptotic consistency for pseudo-likelihood. The proof gives significant insight into contrastive objectives, suggesting that they enforce (soft) probability-ratio constraints between pairs of instantiations. Based on this insight, we propose Contrastive Constraint Generation (CCG),an iterative constraint-generation style algorithm that allows us to learn a log-linear model using only MAP inference. We evaluate

CCG on a scene classification task, showing that it significantly outperforms pseudo-likelihood,contrastive divergence, and a well-known margin-based method.
D. Vickrey, C. Chiung-Yu Lin, D. Koller

Multi-Task Learning of Gaussian Graphical Models (1116)
We present multi-task structure learning for Gaussian graphical models. We discuss uniqueness and boundedness of the optimal solution of the maximization problem. A block coordinate descent method leads to a provably convergent algorithm that generates a sequence of positive definite solutions. Thus, we reduce the original problem into a sequence of strictly convex ℓ_∞ regularized quadratic minimization subproblems. We further show that this subproblem leads to the continuous quadratic knapsack problem, for which very efficient methods exist. Finally, we show promising results in a dataset that captures brain function of cocaine addicted and control subjects under conditions of monetary reward.
J. Honorio, D. Samaras

Learning Programs: A Hierarchical Bayesian Approach (1124)
We are interested in learning programs for multiple related tasks given only a few training examples per task. Since the program for a single task is underdetermined by its data, we introduce a nonparametric hierarchical Bayesian prior over programs which shares statistical strength across multiple tasks. The key challenge is to parametrize this multi-task sharing. For this, we introduce a new representation of programs based on combinatory logic and provide an MCMC algorithm that can perform safe program transformations on this representation to reveal shared inter-program substructures.
P. Liang, M. Jordan, D. Klein

8b: Semi-Supervised Learning 2 (Tamar, Thu 15.40)

Cognitive Models of Test-Item Effects in Human Category Learning (1132)
Imagine two identical people receive exactly the same training on how to classify certain objects. Perhaps surprisingly, we show that one can then manipulate them into classifying some test items in opposite ways, simply depending on what other test items they are asked to classify (without label feedback). We call this the Test-Item Effect, which can be induced by the order or the distribution of test items. We formulate the Test-Item Effect as online semi-supervised learning, and extend three standard human category learning models to explain it.
X. Zhu, B. Gibson, K.-S. Jun, T. Rogers, J. Harrison, Chuck Kalish

A New Analysis of Co-Training (1140)
In this paper, we present a new analysis on co-training, a representative paradigm of disagreement-based semi-supervised learning methods. In our analysis the co-training process is viewed as a combinative label propagation over two views; this provides possibility to bring the graph-based and disagreement-based semi-supervised methods into a unified framework. With the analysis we get some insight that has not been disclosed by previous theoretical studies. In particular, we provide the tex tit {sufficient and necessary} condition for co-training to succeed. We also discuss the relationship to previous theoretical results and give some other interesting implications of our results, such as combination of weight matrices and view split.
W. Wang, Z.-H. Zhou

Learning from Noisy Side Information by Generalized Maximum Entropy Model (1148)
We consider the problem of learning from noisy side information in the form of pairwise constraints. Although many algorithms have been developed to learn from side information, most of them assume perfect pairwise constraints. Given the pairwise constraints are often extracted from data sources such as paper citations, they tend to be noisy and inaccurate. In this paper, we introduce the generalization of maximum entropy model and propose a framework for learning from noisy side information based on the generalized maximum entropy model. The theoretic analysis shows that under certain assumption, the classification model trained from the noisy side information can be very close to the one trained from the perfect side information. Extensive empirical studies verify the effectiveness of the proposed framework.
T. Yang, R. Jin, A. Jain

SVM Classifier Estimation from Group Probabilities (1156)
A learning problem that has only recently gained attention in the machine learning community is that of learning a classifier from group probabilities. It is a learning task that lies somewhere between the well-known tasks of supervised and unsupervised learning, in the sense that for a set of observations we do not know the labels, but for some groups of observations, the frequency distribution of the label is known. This learning problem has important practical applications, for example in privacy-preserving data mining. This paper presents an approach to learn a classifier from group probabilities based on support vector regression and the idea of inverting a classifier calibration process. A detailed analysis will show that this new approach outperforms existing approaches.
S. Rueping

8c: Time-Series Analysis (Rimon, Thu 15.40)

Application of Machine Learning To Epileptic Seizure Detection (1164)
We present and evaluate a machine learning approach to constructing patient-specific classifiers that detect the onset of an epileptic seizure through analysis of the scalp EEG, a non-invasive measure of the brain's electrical activity. This problem is challenging because the brain's electrical activity is composed of numerous classes with overlapping characteristics. The key steps involved in realizing a high performance algorithm included shaping the problem into an appropriate machine learning framework, and identifying the features critical to separating seizure from other types of brain activity. When trained on 2 or more seizures per patient and tested on 916 hours of continuous EEG from 24 patients, our algorithm detected 96\% of 173 test seizures with a median detection delay of 3 seconds and a median false detection rate of 2 false detections per 24 hour period. We also provide information about how to download the CHB-MIT database, which contains the data used in this study.
A. Shoeb, J. Guttag

Dynamical Products of Experts for Modeling Financial Time Series (1172)
Predicting the ``Value at Risk'' of a portfolio of stocks is of great significance in quantitative finance. We introduce a new class models, ``dynamical products of experts'' that treats the latent process over volatilities as an inverse Gamma process. We show that our multivariate volatility models significantly outperform all related Garch and stochastic volatility models which are in popular use in the quantitative finance community.
Y. Chen, Max Welling

Gaussian Process Change Point Models (1180)
We combine Bayesian online change point detection with Gaussian processes to create a nonparametric time series model which can handle change points.The model can be used to locate change points in an online manner;and, unlike other Bayesian online change point detection algorithms, is applicable when temporal correlations in a regime are expected.We show three variations on how to apply Gaussian processes in the change point context, each with their own advantages.We present methods to reduce the computational burden of these models and demonstrate it on several real world data sets.
Y. Saatci, R. Turner, C. Rasmussen

Learning the Linear Dynamical System with ASOS (1188)
We develop a new algorithm, based on EM, for learning the Linear Dynamical System model. Called the method of Approximated Second-Order Statistics (ASOS) our approach achieves dramatically superior computational performance over standard EM through its use of approximations, which we justify with both intuitive explanations and rigorous convergence results. In particular, after an inexpensive pre-computation phase, the iterations of ASOS can be performed in time independent of the length of the training dataset.
J. Martens

8d: Online and Active Learning (Hadas, Thu 15.40)

Budgeted Distribution Learning of Belief Net Parameters (1196)
Most learning algorithms assume that a data set is given initially. We address the common situation where

data is not available initially, but can be obtained, at a cost. We focus on learning Bayesian belief networks (BNs) over discrete variables. As such BNs are models of probabilistic distributions, we consider the generative challenge of learning the parameters, for a fixed structure, that best match the true distribution. We focus on the budgeted learning setting, where there is a known fixed cost c_i for acquiring the value of the i-th feature for any specified instance, and a known total cost to spend acquiring all information. After formally defining this problem from a Bayesian perspective, we first consider allocation algorithms that must decide, before seeing any results, which features of which instances to probe. We show this is NP-hard, even if all variables are independent, then prove that the greedy allocation algorithm IGA is optimal when the costs are uniform and the features are independent, but can otherwise be sub-optimal. We then show that general (non-allocation) policies perform better, and explore the challenges of learning the parameters for general belief networks in this setting, describing conditions for when the obvious round-robin algorithm will, versus will not work optimally. We also explore the effectiveness of this and various other heuristic algorithms.
B. Poczos, L. Li, C. Szepesvari, R. Greiner

Interactive Submodular Set Cover (1204)
We introduce a natural generalization of submodular set cover and exact active learning with a finite hypothesis class (query learning). We call this new problem interactive submodular set cover. Applications include advertising in social networks with hidden information. We give an approximation guarantee for a novel greedy algorithm and give a hardness of approximation result which matches up to constant factors. We also discuss negative results for simpler approaches and present encouraging early experimental results.
A. Guillory, J. Bilmes

Learning optimally diverse rankings over large document collections (1212)
Most learning to rank research has assumed that the utility of different documents is independent, which results in learned ranking functions that return redundant results. The few approaches that avoid this have rather un satisfyingly lacked theoretical foundations, or do not scale. We present a learning-to-rank formulation that optimizes the fraction of satisfied users, with a scalable algorithm that explicitly takes document similarity and ranking context into account. We present theoretical justifications for this approach, as well as a near-optimal algorithm. Our evaluation adds optimizations that improve empirical performance, and shows that our algorithms learn orders of magnitude more quickly than previous approaches.
A. Slivkins, F. Radlinski, S. Gollapudi

Budgeted Nonparametric Learning from Data Streams (1220)
We consider the problem of extracting informative exemplars from a data stream. Examples of this problem include exemplar-based clustering and nonparametric inference such as Gaussian process regression on massive data sets. We show that these problems require maximization of a submodular function that captures the informativeness of a set of exemplars, over a data stream. We develop an efficient algorithm, Stream Greedy, which is guaranteed to obtain a constant fraction of the value achieved by the optimal solution to this NP-hard optimization problem. We extensively evaluate our algorithm on large real- world data sets.
R. Gomes, A. Krause

8e: Multi-Label and Multi-Instance Learning (Arava, Thu 15.40)

A Conditional Random Field for Multi-Instance Learning (1228)
We present MI-CRF, a conditional random field (CRF) model for multiple instance learning (MIL). MI-CRF models bags as nodes in a CRF with instances as their states. It combines discriminative unary instance classifiers and pairwise dissimilarity measures. We show that both forces improve the classification performance. Unlike other approaches, MI-CRF considers all bags jointly during training as well as during testing. This makes it possible to classify test bags in an imputation setup. The parameters of MI-CRF are learned using constraint generation. Furthermore, we show that MI-CRF can incorporate previous MIL algorithms to improve on their results. MI-CRF obtains competitive results on five standard MIL datasets.
T. Deselaers, V. Ferrari

Large Scale Max-Margin Multi-Label Classification with Priors (1236)
We propose a max-margin formulation for the multi-label classification problem where the goal is to tag a data point with a set of pre-specified labels. Given a set of L labels, a data point can be tagged with any of the 2^L possible subsets. The main challenge therefore lies in optimizing over this exponentially large label space subject to label correlations .Existing solutions take either of two approaches. The first assumes, it a priori , that there are no label correlations and independently trains a classifier for each label (as is done in the 1-vs-Allheuristic). This reduces the problem complexity from exponential to linear and such methods can scale to large problems. The second approach explicitly models correlations by pair wise label interactions. However, the complexity remains exponential unless one assumes that label correlations are sparse. Furthermore, the learnt correlations reflect the training set biases.We take a middle approach that assumes labels are correlated but does not incorporate pair wise label terms in the prediction function. We show that the complexity can still be reduced from exponential to linear while modeling dense pair wise label correlations. By incorporating correlation priors we can overcome training set biases and improve prediction accuracy. We provide a principled interpretation of the 1-vs-All method and show that it arises as a special case of our formulation. We also develop efficient optimization algorithms that can be orders of magnitude faster than the state-of-the-art.
B. Hariharan, Lihi Zelnik-Manor, S.V.N. Vishwanathan, M. Varma

Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains (1244)
In the realm of multilabel classification (MLC), it has become an opinio communis that optimal predictive performance can only be achieved by learners that explicitly take label dependence into account. The goal of this paper is to elaborate on this postulate in a critical way. To this end, we formalize and analyze MLC within a probabilistic setting. Thus, it becomes possible to look at the problem from the point of view of risk minimization and Bayes optimal prediction. Moreover, inspired by our probabilistic setting, we propose a new method for MLC that generalizes and outperforms another approach, called classifier chains, that was recently introduced in the literature.
K. Dembczynski, W. Cheng, E. Hullermeier

Graded Multilabel Classification: The Ordinal Case (1252)
We propose a generalization of multilabel classification that we refer to as graded multilabel classification. The key idea is that, instead of requesting a yes-no answer to the question of class membership or, say, relevance of a class label for an instance, we allow for a graded membership of an instance, measured on an ordinal scale of membership degrees. This extension is motivated by practical applications in which a graded or partial class membership is natural. Apart from introducing the basic setting, we propose two general strategies for reducing graded multilabel problems to conventional (multilabel) classification problems. Moreover, we address the question of how to extend performance metrics commonly used in multilabel classification to the graded setting, and present first experimental results.
W. Cheng, K. Dembczynski, E. Hullermeier