Agnostic Active Learning Without Constraints
Alina Beygelzimer, Daniel J. Hsu, John Langford, Tong Zhang
We present and analyze an agnostic active learning algorithm that works without keeping a version space. This is unlike all previous approaches where a restricted set of candidate hypotheses is maintained throughout learning, and only hypotheses from this set are ever returned. By avoiding this version space approach, our algorithm sheds the computational burden and brittleness associated with maintaining version spaces, yet still allows for substantial improvements over supervised learning for classification.
****
A Dirty Model for Multi-task Learning
Ali Jalali, Sujay Sanghavi, Chao Ruan, Pradeep Ravikumar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Generative Local Metric Learning for Nearest Neighbor Classification
Yung-kyun Noh, Byoung-tak Zhang, Daniel Lee
We consider the problem of learning a local metric to enhance the performance of nearest neighbor classification. Conventional metric learning methods attempt to separate data distributions in a purely discriminative manner; here we show how to take advantage of information from parametric generative models. We focus on the bias in the information-theoretic error arising from finite sampling effects, and find an appropriate local metric that maximally reduces the bias based upon knowledge from generative models. As a byproduct, the asymptotic theoretical analysis in this work relates metric learning with dimensionality reduction, which was not understood from previous discriminative approaches. Empirical experiments show that this learned local metric enhances the discriminative nearest neighbor performance on various datasets using simple class conditional generative models.
****
Relaxed Clipping: A Global Training Method for Robust Regression and Classification
Min Yang, Linli Xu, Martha White, Dale Schuurmans, Yao-liang Yu
Robust regression and classification are often thought to require non-convex loss functions that prevent scalable, global training. However, such a view neglects the possibility of reformulated training methods that can yield practically solvable alternatives. A natural way to make a loss function more robust to outliers is to truncate loss values that exceed a maximum threshold. We demonstrate that a relaxation of this form of ``loss clipping'' can be made globally solvable and applicable to any standard loss while guaranteeing robustness against outliers. We present a generic procedure that can be applied to standard loss functions and demonstrate improved robustness in regression and classification problems.
****
Linear readout from a neural population with partial correlation data
Adrien Wohrer, Ranulfo Romo, Christian K. Machens
How much information does a neural population convey about a stimulus? Answers to this question are known to strongly depend on the correlation of response variability in neural populations. These noise correlations, however, are essentially immeasurable as the number of parameters in a noise correlation matrix grows quadratically with population size. Here, we suggest to bypass this problem by imposing a parametric model on a noise correlation matrix. Our basic assumption is that noise correlations arise due to common inputs between neurons. On average, noise correlations will therefore reflect signal correlations, which can be measured in neural populations. We suggest an explicit parametric dependency between signal and noise correlations. We show how this dependency can be used to fill the gaps" in noise correlations matrices using an iterative application of the Wishart distribution over positive definitive matrices. We apply our method to data from the primary somatosensory cortex of monkeys performing a two-alternativ

e-forced choice task. We compare the discrimination thresholds read out from the population of recorded neurons with the discrimination threshold of the monkey and show that our method predicts different results than simpler, average schemes of noise correlations."

****

On Herding and the Perceptron Cycling Theorem

Andrew Gelfand, Yutian Chen, Laurens Maaten, Max Welling

The paper develops a connection between traditional perceptron algorithms and recently introduced herding algorithms. It is shown that both algorithms can be viewed as an application of the perceptron cycling theorem. This connection strengthens some herding results and suggests new (supervised) herding algorithms that, like CRFs or discriminative RBMs, make predictions by conditioning on the input attributes. We develop and investigate variants of conditional herding, and show that conditional herding leads to practical algorithms that perform better than or on par with related classifiers such as the voted perceptron and the discriminative RBM.

****

Tiled convolutional neural networks

Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pang Koh, Quoc Le, Andrew Ng

Convolutional neural networks (CNNs) have been successfully applied to many tasks such as digit and object recognition. Using convolutional (tied) weights significantly reduces the number of parameters that have to be learned, and also allows translational invariance to be hard-coded into the architecture. In this paper, we consider the problem of learning invariances, rather than relying on hard-coding. We propose tiled convolution neural networks (Tiled CNNs), which use a regular “tiled” pattern of tied weights that does not require that adjacent hidden units share identical weights, but instead requires only that hidden units k steps away from each other to have tied weights. By pooling over neighboring units, this architecture is able to learn complex invariances (such as scale and rotational invariance) beyond translational invariance. Further, it also enjoys much of CNNs’ advantage of having a relatively small number of learned parameters (such as ease of learning and greater scalability). We provide an efficient learning algorithm for Tiled CNNs based on Topographic ICA, and show that learning complex invariant features allows us to achieve highly competitive results for both the NORB and CIFAR-10 datasets.

****

Decomposing Isotonic Regression for Efficiently Solving Large Problems

Ronny Luss, Saharon Rosset, Moni Shahar

A new algorithm for isotonic regression is presented based on recursively partitioning the solution space. We develop efficient methods for each partitioning subproblem through an equivalent representation as a network flow problem, and prove that this sequence of partitions converges to the global solution. These network flow problems can further be decomposed in order to solve very large problems. Success of isotonic regression in prediction and our algorithm's favorable computational properties are demonstrated through simulated examples as large as 2 x10^5 variables and 10^7 constraints.

****

Learning Kernels with Radiuses of Minimum Enclosing Balls

Kun Gai, Guangyun Chen, Chang-shui Zhang

In this paper, we point out that there exist scaling and initialization problems in most existing multiple kernel learning (MKL) approaches, which employ the large margin principle to jointly learn both a kernel and an SVM classifier. The reason is that the margin itself can not well describe how good a kernel is due to the negligence of the scaling. We use the ratio between the margin and the radius of the minimum enclosing ball to measure the goodness of a kernel, and present a new minimization formulation for kernel learning. This formulation is invariant to scalings of learned kernels, and when learning linear combination of basis kernels it is also invariant to scalings of basis kernels and to the types (e.g., L1 or L2) of norm constraints on combination coefficients. We establish the differentiability of our formulation, and propose a gradient projection algorit

hm for kernel learning. Experiments show that our method significantly outperfor ms both SVM with the uniform combination of basis kernels and other state-of-art MKL approaches.
****

Label Embedding Trees for Large Multi-Class Tasks
Samy Bengio, Jason Weston, David Grangier
Multi-class classification becomes challenging at test time when the number of c lasses is very large and testing against every possible class can become computa tionally infeasible. This problem can be alleviated by imposing (or learning) a structure over the set of classes. We propose an algorithm for learning a tree-s tructure of classifiers which, by optimizing the overall tree loss, provides sup erior accuracy to existing tree labeling methods. We also propose a method that learns to embed labels in a low dimensional space that is faster than non-embedd ing approaches and has superior accuracy to existing embedding approaches. Final ly we combine the two ideas resulting in the label embedding tree that outperfor ms alternative methods including One-vs-Rest while being orders of magnitude fas ter.
****

Deep Coding Network
Yuanqing Lin, Tong Zhang, Shenghuo Zhu, Kai Yu
This paper proposes a principled extension of the traditional single-layer flat sparse coding scheme, where a two-layer coding scheme is derived based on theore tical analysis of nonlinear functional approximation that extends recent results for local coordinate coding. The two-layer approach can be easily generalized t o deeper structures in a hierarchical multiple-layer manner. Empirically, it is shown that the deep coding approach yields improved performance in benchmark dat asets.
****

Transduction with Matrix Completion: Three Birds with One Stone
Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, Jerry Zhu
We pose transductive classification as a matrix completion problem. By assuming the underlying matrix has a low rank, our formulation is able to handle three pr oblems simultaneously: i) multi-label learning, where each item has more than on e label, ii) transduction, where most of these labels are unspecified, and iii) missing data, where a large number of features are missing. We obtained satisfac tory results on several real-world tasks, suggesting that the low rank assumptio n may not be as restrictive as it seems. Our method allows for different loss fu nctions to apply on the feature and label entries of the matrix. The resulting n uclear norm minimization problem is solved with a modified fixed-point continuat ion method that is guaranteed to find the global optimum.
****

Extended Bayesian Information Criteria for Gaussian Graphical Models
Rina Foygel, Mathias Drton
Gaussian graphical models with sparsity in the inverse covariance matrix are of significant interest in many modern applications. For the problem of recovering the graphical structure, information criteria provide useful optimization object ives for algorithms searching through sets of graphs or for selection of tuning parameters of other methods such as the graphical lasso, which is a likelihood p enalization technique. In this paper we establish the asymptotic consistency of an extended Bayesian information criterion for Gaussian graphical models in a sc enario where both the number of variables p and the sample size n grow. Compared to earlier work on the regression case, our treatment allows for growth in the number of non-zero parameters in the true model, which is necessary in order to cover connected graphs. We demonstrate the performance of this criterion on simu lated data when used in conjuction with the graphical lasso, and verify that the criterion indeed performs better than either cross-validation or the ordinary B ayesian information criterion when p and the number of non-zero parameters q bot h scale with n.
****

Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and

Surfaces
Abhinav Gupta, Martial Hebert, Takeo Kanade, David Blei
There has been a recent push in extraction of 3D spatial layout of scenes. Howev
er, none of these approaches model the 3D interaction between objects and the sp
atial layout. In this paper, we argue for a parametric representation of objects
 in 3D, which allows us to incorporate volumetric constraints of the physical wo
rld. We show that augmenting current structured prediction techniques with volum
etric reasoning significantly improves the performance of the state-of-the-art.
****
A Computational Decision Theory for Interactive Assistants
Alan Fern, Prasad Tadepalli
We study several classes of interactive assistants from the points of view of de
cision theory and computational complexity. We first introduce a class of POMDPs
 called hidden-goal MDPs (HGMDPs), which formalize the problem of interactively
assisting an agent whose goal is hidden and whose actions are observable. In spi
te of its restricted nature, we show that optimal action selection in finite hor
izon HGMDPs is PSPACE-complete even in domains with deterministic dynamics. We t
hen introduce a more restricted model called helper action MDPs (HAMDPs), where
the assistant's action is accepted by the agent when it is helpful, and can be e
asily ignored by the agent otherwise. We show classes of HAMDPs that are complet
e for PSPACE and NP along with a polynomial time class. Furthermore, we show tha
t for general HAMDPs a simple myopic policy achieves a regret, compared to an om
niscient assistant, that is bounded by the entropy of the initial goal distribut
ion. A variation of this policy is shown to achieve worst-case regret that is lo
garithmic in the number of goals for any goal distribution.
****
Large Margin Multi-Task Metric Learning
Shibin Parameswaran, Kilian Q. Weinberger
Multi-task learning (MTL) improves the prediction performance on multiple, diffe
rent but related, learning problems through shared parameters or representations
. One of the most prominent multi-task learning algorithms is an extension to sv
ms by Evgeniou et al. Although very elegant, multi-task svm is inherently restri
cted by the fact that support vector machines require each class to be addressed
 explicitly with its own weight vector which, in a multi-task setting, requires
the different learning tasks to share the same set of classes. This paper propos
es an alternative formulation for multi-task learning by extending the recently
published large margin nearest neighbor (lmnn) algorithm to the MTL paradigm. In
stead of relying on separating hyperplanes, its decision function is based on th
e nearest neighbor rule which inherently extends to many classes and becomes a n
atural fit for multitask learning. We evaluate the resulting multi-task lmnn on
real-world insurance data and speech classification problems and show that it co
nsistently outperforms single-task kNN under several metrics and state-of-the-ar
t MTL classifiers.
****
Evaluating neuronal codes for inference using Fisher information
Haefner Ralf, Matthias Bethge
Many studies have explored the impact of response variability on the quality of
sensory codes. The source of this variability is almost always assumed to be int
rinsic to the brain. However, when inferring a particular stimulus property, var
iability associated with other stimulus attributes also effectively act as noise
. Here we study the impact of such stimulus-induced response variability for the
 case of binocular disparity inference. We characterize the response distributio
n for the binocular energy model in response to random dot stereograms and find
it to be very different from the Poisson-like noise usually assumed. We then com
pute the Fisher information with respect to binocular disparity, present in the
monocular inputs to the standard model of early binocular processing, and thereb
y obtain an upper bound on how much information a model could theoretically extr
act from them. Then we analyze the information loss incurred by the different wa
ys of combining those inputs to produce a scalar single-neuron response. We find
 that in the case of depth inference, monocular stimulus variability places a gr

eater limit on the extractable information than intrinsic neuronal noise for typ
ical spike counts. Furthermore, the largest loss of information is incurred by t
he standard model for position disparity neurons (tuned-excitatory), that are th
e most ubiquitous in monkey primary visual cortex, while more information from t
he inputs is preserved in phase-disparity neurons (tuned-near or tuned-far) prim
arily found in higher cortical regions.
****

Guaranteed Rank Minimization via Singular Value Projection
Prateek Jain, Raghu Meka, Inderjit Dhillon
Minimizing the rank of a matrix subject to affine constraints is a fundamental p
roblem with many important applications in machine learning and statistics. In t
his paper we propose a simple and fast algorithm SVP (Singular Value Projection)
 for rank minimization under affine constraints ARMP and show that SVP recovers
the minimum rank solution for affine constraints that satisfy a Restricted Isome
try Property} (RIP). Our method guarantees geometric convergence rate even in th
e presence of noise and requires strictly weaker assumptions on the RIP constant
s than the existing methods. We also introduce a Newton-step for our SVP framewo
rk to speed-up the convergence with substantial empirical gains. Next, we addres
s a practically important application of ARMP - the problem of low-rank matrix c
ompletion, for which the defining affine constraints do not directly obey RIP, h
ence the guarantees of SVP do not hold. However, we provide partial progress tow
ards a proof of exact recovery for our algorithm by showing a more restricted is
ometry property and observe empirically that our algorithm recovers low-rank Inc
oherent matrices from an almost optimal number of uniformly sampled entries. We
also demonstrate empirically that our algorithms outperform existing methods, su
ch as those of \cite{CaiCS2008,LeeB2009b, KeshavanOM2009}, for ARMP and the matr
ix completion problem by an order of magnitude and are also more robust to noise
 and sampling schemes. In particular, results show that our SVP-Newton method is
 significantly robust to noise and performs impressively on a more realistic pow
er-law sampling scheme for the matrix completion problem.
****

Double Q-learning
Hado Hasselt
In some stochastic environments the well-known reinforcement learning algorithm
Q-learning performs very poorly. This poor performance is caused by large overes
timations of action values. These overestimations result from a positive bias th
at is introduced because Q-learning uses the maximum action value as an approxim
ation for the maximum expected action value. We introduce an alternative way to
approximate the maximum expected value for any set of random variables. The obta
ined double estimator method is shown to sometimes underestimate rather than ove
restimate the maximum expected value. We apply the double estimator to Q-learnin
g to construct Double Q-learning, a new off-policy reinforcement learning algori
thm. We show the new algorithm converges to the optimal policy and that it perfo
rms well in some settings in which Q-learning performs poorly due to its overest
imation.
****

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

Efficient and Robust Feature Selection via Joint ■2,1-Norms Minimization
Feiping Nie, Heng Huang, Xiao Cai, Chris Ding
Feature selection is an important component of many machine learning application
s. Especially in many bioinformatics tasks, efficient and robust feature selecti
on methods are desired to extract meaningful features and eliminate noisy ones.
In this paper, we propose a new robust feature selection method with emphasizing
 joint ■2,1-norm minimization on both loss function and regularization. The ■2,1

-norm based loss function is robust to outliers in data points and the ■2,1-norm regularization selects features across all data points with joint sparsity. An efficient algorithm is introduced with proved convergence. Our regression based objective makes the feature selection process more efficient. Our method has bee n applied into both genomic and proteomic biomarkers discovery. Extensive empiri cal studies were performed on six data sets to demonstrate the effectiveness of our feature selection method.

****

Repeated Games against Budgeted Adversaries

Jacob D. Abernethy, Manfred K. K. Warmuth

We study repeated zero-sum games against an adversary on a budget. Given that an adversary has some constraint on the sequence of actions that he plays, we cons ider what ought to be the player's best mixed strategy with knowledge of this bu dget. We show that, for a general class of normal-form games, the minimax strate gy is indeed efficiently computable and relies on a random playout" technique. W e give three diverse applications of this algorithmic template: a cost-sensitive "Hedge" setting, a particular problem in Metrical Task Systems, and the design of combinatorial prediction markets."

****

Switching state space model for simultaneously estimating state transitions and nonstationary firing rates

Ken Takiyama, Masato Okada

We propose an algorithm for simultaneously estimating state transitions among ne ural states, the number of neural states, and nonstationary firing rates using a switching state space model (SSSM). This model enables us to detect state tran sitions based not only on the discontinuous changes of mean firing rates but als o on discontinuous changes in temporal profiles of firing rates, e.g., temporal correlation. We derive a variational Bayes algorithm for a non-Gaussian SSSM wh ose non-Gaussian property is caused by binary spike events. Synthetic data anal ysis reveals the high performance of our algorithm in estimating state transitio ns, the number of neural states, and nonstationary firing rates compared to prev ious methods. We also analyze neural data recorded from the medial temporal are a. The statistically detected neural states probably coincide with transient an d sustained states, which have been detected heuristically. Estimated parameters suggest that our algorithm detects the state transition based on discontinuous change in the temporal correlation of firing rates, which transitions previous m ethods cannot detect. This result suggests the advantage of our algorithm in re al-data analysis.

****

Why are some word orders more common than others? A uniform information density account

Luke Maurits, Dan Navarro, Amy Perfors

Languages vary widely in many ways, including their canonical word order. A basi c aspect of the observed variation is the fact that some word orders are much mo re common than others. Although this regularity has been recognized for some tim e, it has not been well-explained. In this paper we offer an information-theoret ic explanation for the observed word-order distribution across languages, based on the concept of Uniform Information Density (UID). We suggest that object-firs t languages are particularly disfavored because they are highly non-optimal if t he goal is to distribute information content approximately evenly throughout a s entence, and that the rest of the observed word-order distribution is at least p artially explainable in terms of UID. We support our theoretical analysis with d ata from child-directed speech and experimental work.

****

Getting lost in space: Large sample analysis of the resistance distance

Ulrike Luxburg, Agnes Radl, Matthias Hein

The commute distance between two vertices in a graph is the expected time it t akes a random walk to travel from the first to the second vertex and back. We study the behavior of the commute distance as the size of the underlying graph increases. We prove that the commute distance converges to an expression that

does not take into account the structure of the graph at all and that is completely meaningless as a distance function on the graph. Consequently, the use of the raw commute distance for machine learning purposes is strongly discouraged for large graphs and in high dimensions. As an alternative we introduce the amplified commute distance that corrects for the undesired large sample effects.
****

Multiparty Differential Privacy via Aggregation of Locally Trained Classifiers
Manas Pathak, Shantanu Rane, Bhiksha Raj
As increasing amounts of sensitive personal information finds its way into data repositories, it is important to develop analysis mechanisms that can derive aggregate information from these repositories without revealing information about individual data instances. Though the differential privacy model provides a framework to analyze such mechanisms for databases belonging to a single party, this framework has not yet been considered in a multi-party setting. In this paper, we propose a privacy-preserving protocol for composing a differentially private aggregate classifier using classifiers trained locally by separate mutually untrusting parties. The protocol allows these parties to interact with an untrusted curator to construct additive shares of a perturbed aggregate classifier. We also present a detailed theoretical analysis containing a proof of differential privacy of the perturbed aggregate classifier and a bound on the excess risk introduced by the perturbation. We verify the bound with an experimental evaluation on a real dataset.
****

Convex Multiple-Instance Learning by Estimating Likelihood Ratio
Fuxin Li, Cristian Sminchisescu
Multiple-Instance learning has been long known as a hard non-convex problem. In this work, we propose an approach that recasts it as a convex likelihood ratio estimation problem. Firstly, the constraint in multiple-instance learning is reformulated into a convex constraint on the likelihood ratio. Then we show that a joint estimation of a likelihood ratio function and the likelihood on training instances can be learned convexly. Theoretically, we prove a quantitative relationship between the risk estimated under the 0-1 classification loss, and under a loss function for likelihood ratio estimation. It is shown that our likelihood ratio estimation is generally a good surrogate for the 0-1 loss, and separates positive and negative instances well. However with the joint estimation it tends to underestimate the likelihood of an example to be positive. We propose to use these likelihood ratio estimates as features, and learn a linear combination on them to classify the bags. Experiments on synthetic and real datasets show the superiority of the approach.
****

Short-term memory in neuronal networks through dynamical compressed sensing
Surya Ganguli, Haim Sompolinsky
Recent proposals suggest that large, generic neuronal networks could store memory traces of past input sequences in their instantaneous state. Such a proposal raises important theoretical questions about the duration of these memory traces and their dependence on network size, connectivity and signal statistics. Prior work, in the case of gaussian input sequences and linear neuronal networks, shows that the duration of memory traces in a network cannot exceed the number of neurons (in units of the neuronal time constant), and that no network can out-perform an equivalent feedforward network. However a more ethologically relevant scenario is that of sparse input sequences. In this scenario, we show how linear ne

ural networks can essentially perform compressed sensing (CS) of past inputs, th
ereby attaining a memory capacity that {\it exceeds} the number of neurons. This
 enhanced capacity is achieved by a class of ``orthogonal recurrent networks and
 not by feedforward networks or generic recurrent networks. We exploit technique
s from the statistical physics of disordered systems to analytically compute the
 decay of memory traces in such networks as a function of network size, signal s
parsity and integration time. Alternately, viewed purely from the perspective of
 CS, this work introduces a new ensemble of measurement matrices derived from dy
namical systems, and provides a theoretical analysis of their asymptotic perform
ance."
****

The Multidimensional Wisdom of Crowds
Peter Welinder, Steve Branson, Pietro Perona, Serge Belongie
Distributing labeling tasks among hundreds or thousands of annotators is an incr
easingly important method for annotating large datasets. We present a method for
 estimating the underlying value (e.g. the class) of each image from (noisy) ann
otations provided by multiple annotators. Our method is based on a model of the
image formation and annotation process. Each image has different characteristics
 that are represented in an abstract Euclidean space. Each annotator is modeled
as a multidimensional entity with variables representing competence, expertise a
nd bias. This allows the model to discover and represent groups of annotators th
at have different sets of skills and knowledge, as well as groups of images that
 differ qualitatively. We find that our model predicts ground truth labels on bo
th synthetic and real data more accurately than state of the art methods. Experi
ments also show that our model, starting from a set of binary labels, may discov
er rich information, such as different "schools of thought" amongst the annotato
rs, and can group together images belonging to separate categories.
****

Boosting Classifier Cascades
Nuno Vasconcelos, Mohammad Saberian
The problem of optimal and automatic design of a detector cascade is considered.
 A novel mathematical model is introduced for a cascaded detector. This model is
 analytically tractable, leads to recursive computation, and accounts for both c
lassification and complexity. A boosting algorithm, FCBoost, is proposed for ful
ly automated cascade design. It exploits the new cascade model, minimizes a Lagr
angian cost that accounts for both classification risk and complexity. It search
es the space of cascade configurations to automatically determine the optimal nu
mber of stages and their predictors, and is compatible with bootstrapping of neg
ative examples and cost sensitive learning. Experiments show that the resulting
cascades have state-of-the-art performance in various computer vision problems.
****

Implicitly Constrained Gaussian Process Regression for Monocular Non-Rigid Pose
Estimation
Mathieu Salzmann, Raquel Urtasun
Estimating 3D pose from monocular images is a highly ambiguous problem. Physical
 constraints can be exploited to restrict the space of feasible configurations.
In this paper we propose an approach to constraining the prediction of a discrim
inative predictor. We first show that the mean prediction of a Gaussian process
implicitly satisfies linear constraints if those constraints are satisfied by th
e training examples. We then show how, by performing a change of variables, a GP
 can be forced to satisfy quadratic constraints. As evidenced by the experiments
, our method outperforms state-of-the-art approaches on the tasks of rigid and n
on-rigid pose estimation.
****

Effects of Synaptic Weight Diffusion on Learning in Decision Making Networks
Kentaro Katahira, Kazuo Okanoya, Masato Okada
When animals repeatedly choose actions from multiple alternatives, they can allo
cate their choices stochastically depending on past actions and outcomes. It is
commonly assumed that this ability is achieved by modifications in synaptic weig
hts related to decision making. Choice behavior has been empirically found to fo

llow Herrnstein’s matching law. Loewenstein & Seung (2006) demonstrated that mat ching behavior is a steady state of learning in neural networks if the synaptic weights change proportionally to the covariance between reward and neural activi ties. However, their proof did not take into account the change in entire synapt ic distributions. In this study, we show that matching behavior is not necessari ly a steady state of the covariance-based learning rule when the synaptic streng th is sufficiently strong so that the fluctuations in input from individual sens ory neurons influence the net input to output neurons. This is caused by the inc reasing variance in the input potential due to the diffusion of synaptic weights . This effect causes an undermatching phenomenon, which has been observed in man y behavioral experiments. We suggest that the synaptic diffusion effects provide a robust neural mechanism for stochastic choice behavior.

****

Interval Estimation for Reinforcement-Learning Algorithms in Continuous-State Do mains

Martha White, Adam White

The reinforcement learning community has explored many approaches to obtain- ing value estimates and models to guide decision making; these approaches, how- eve r, do not usually provide a measure of confidence in the estimate. Accurate esti mates of an agent’s confidence are useful for many applications, such as bi- asi ng exploration and automatically adjusting parameters to reduce dependence on pa rameter-tuning. Computing confidence intervals on reinforcement learning value e stimates, however, is challenging because data generated by the agent- environme nt interaction rarely satisfies traditional assumptions. Samples of value- estim ates are dependent, likely non-normally distributed and often limited, partic- u larly in early learning when confidence estimates are pivotal. In this work, we investigate how to compute robust confidences for value estimates in continuous Markov decision processes. We illustrate how to use bootstrapping to compute con fidence intervals online under a changing policy (previously not possible) and p rove validity under a few reasonable assumptions. We demonstrate the applica- bi lity of our confidence estimation algorithms with experiments on exploration, pa rameter estimation and tracking.

****

Object Bank: A High-Level Image Representation for Scene Classification & Semant ic Feature Sparsification

Li-jia Li, Hao Su, Li Fei-fei, Eric Xing

Robust low-level image features have been proven to be effective representations for a variety of visual recognition tasks such as object recognition and scene classification; but pixels, or even local image patches, carry little semantic m eanings. For high level visual tasks, such low-level image representations are p otentially not enough. In this paper, we propose a high-level image representati on, called the Object Bank, where an image is represented as a scale invariant r esponse map of a large number of pre-trained generic object detectors, blind to the testing dataset or visual task. Leveraging on the Object Bank representation , superior performances on high level visual recognition tasks can be achieved w ith simple off-the-shelf classifiers such as logistic regression and linear SVM. Sparsity algorithms make our representation more efficient and scalable for lar ge scene datasets, and reveal semantically meaningful feature patterns.

****

Using body-anchored priors for identifying actions in single images

Leonid Karlinsky, Michael Dinerstein, Shimon Ullman

This paper presents an approach to the visual recognition of human actions using only single images as input. The task is easy for humans but difficult for curr ent approaches to object recognition, because action instances may be similar in terms of body pose, and often require detailed examination of relations between participating objects and body parts in order to be recognized. The proposed ap proach applies a two-stage interpretation procedure to each training and test im age. The first stage produces accurate detection of the relevant body parts of t he actor, forming a prior for the local evidence needed to be considered for ide ntifying the action. The second stage extracts features that are ‘anchored’ to t

he detected body parts, and uses these features and their feature-to-part relati ons in order to recognize the action. The body anchored priors we propose apply to a large range of human actions. These priors allow focusing on the relevant r egions and relations, thereby significantly simplifying the learning process and increasing recognition performance.

****

Reward Design via Online Gradient Ascent

Jonathan Sorg, Richard L. Lewis, Satinder Singh

Recent work has demonstrated that when artificial agents are limited in their ab ility to achieve their goals, the agent designer can benefit by making the agent 's goals different from the designer's. This gives rise to the optimization prob lem of designing the artificial agent's goals---in the RL framework, designing t he agent's reward function. Existing attempts at solving this optimal reward pro blem do not leverage experience gained online during the agent's lifetime nor do they take advantage of knowledge about the agent's structure. In this work, we develop a gradient ascent approach with formal convergence guarantees for approx imately solving the optimal reward problem online during an agent's lifetime. We show that our method generalizes a standard policy gradient approach, and we de monstrate its ability to improve reward functions in agents with various forms o f limitations.

****

Universal Consistency of Multi-Class Support Vector Classification

Tobias Glasmachers

Steinwart was the ■rst to prove universal consistency of support vector machine classi■cation. His proof analyzed the ‘standard’ support vector machine classi■ er, which is restricted to binary classi■cation problems. In contrast, recent a nalysis has resulted in the common belief that several extensions of SVM classi ■cation to more than two classes are inconsistent. Countering this belief, we proof the universal consistency of the multi-class support vector machine by Cra mmer and Singer. Our proof extends Steinwart’s techniques to the multi-class cas e.

****

Supervised Clustering

Pranjal Awasthi, Reza Zadeh

Despite the ubiquity of clustering as a tool in unsupervised learning, there is not yet a consensus on a formal theory, and the vast majority of work in this d irection has focused on unsupervised clustering. We study a recently proposed fr amework for supervised clustering where there is access to a teacher. We give an improved generic algorithm to cluster any concept class in that model. Our alg orithm is query-efficient in the sense that it involves only a small amount of interaction with the teacher. We also present and study two natural generalizati ons of the model. The model assumes that the teacher response to the algorithm is perfect. We eliminate this limitation by proposing a noisy model and give an algorithm for clustering the class of intervals in this noisy model. We also p ropose a dynamic model where the teacher sees a random subset of the points. Fi nally, for datasets satisfying a spectrum of weak to strong properties, we give query bounds, and show that a class of clustering functions containing Single-Linkage will find the target clustering under the strongest property.

****

Inferring Stimulus Selectivity from the Spatial Structure of Neural Network Dyna mics

Kanaka Rajan, L Abbott, Haim Sompolinsky

How are the spatial patterns of spontaneous and evoked population responses rela ted? We study the impact of connectivity on the spatial pattern of fluctuations in the input-generated response of a neural network, by comparing the distributi on of evoked and intrinsically generated activity across the different units. We develop a complementary approach to principal component analysis in which separ ate high-variance directions are typically derived for each input condition. We analyze subspace angles to compute the difference between the shapes of trajecto ries corresponding to different network states, and the orientation of the low-d

imensional subspaces that driven trajectories occupy within the full space of ne uronal activity. In addition to revealing how the spatiotemporal structure of sp ontaneous activity affects input-evoked responses, these methods can be used to infer input selectivity induced by network dynamics from experimentally accessib le measures of spontaneous activity (e.g. from voltage- or calcium-sensitive opt ical imaging experiments). We conclude that the absence of a detailed spatial ma p of afferent inputs and cortical connectivity does not limit our ability to des ign spatially extended stimuli that evoke strong responses.
****

Distributionally Robust Markov Decision Processes
Huan Xu, Shie Mannor
We consider Markov decision processes where the values of the parameters are unc ertain. This uncertainty is described by a sequence of nested sets (that is, eac h set contains the previous one), each of which corresponds to a probabilistic g uarantee for a different confidence level so that a set of admissible probabilit y distributions of the unknown parameters is specified. This formulation models the case where the decision maker is aware of and wants to exploit some (yet imp recise) a-priori information of the distribution of parameters, and arises natur ally in practice where methods to estimate the confidence region of parameters a bound. We propose a decision criterion based on distributional robustness: the o ptimal policy maximizes the expected total reward under the most adversarial pro bability distribution over realizations of the uncertain parameters that is admi ssible (i.e., it agrees with the a-priori information). We show that finding the optimal distributionally robust policy can be reduced to a standard robust MDP where the parameters belong to a single uncertainty set, hence it can be compute d in polynomial time under mild technical conditions.
****

Empirical Risk Minimization with Approximations of Probabilistic Grammars
Noah A. Smith, Shay Cohen
Probabilistic grammars are generative statistical models that are useful for co mpositional and sequential structures. We present a framework, reminiscent of st ructural risk minimization, for empirical risk minimization of the parameters of a fixed probabilistic grammar using the log-loss. We derive sample complexity b ounds in this framework that apply both to the supervised setting and the unsupe rvised setting.
****

MAP Estimation for Graphical Models by Likelihood Maximization
Akshat Kumar, Shlomo Zilberstein
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Identifying graph-structured activation patterns in networks
James Sharpnack, Aarti Singh
We consider the problem of identifying an activation pattern in a complex, large -scale network that is embedded in very noisy measurements. This problem is rele vant to several applications, such as identifying traces of a biochemical spread by a sensor network, expression levels of genes, and anomalous activity or cong estion in the Internet. Extracting such patterns is a challenging task specially if the network is large (pattern is very high-dimensional) and the noise is so excessive that it masks the activity at any single node. However, typically ther e are statistical dependencies in the network activation process that can be lev eraged to fuse the measurements of multiple nodes and enable reliable extraction of high-dimensional noisy patterns. In this paper, we analyze an estimator base d on the graph Laplacian eigenbasis, and establish the limits of mean square err or recovery of noisy patterns arising from a probabilistic (Gaussian or Ising) m odel based on an arbitrary graph structure. We consider both deterministic and p robabilistic network evolution models, and our results indicate that by leveragi ng the network interaction structure, it is possible to consistently recover hig

h-dimensional patterns even when the noise variance increases with network size.
****
Size Matters: Metric Visual Search Constraints from Monocular Metadata
Mario Fritz, Kate Saenko, Trevor Darrell
Metric constraints are known to be highly discriminative for many objects, but i f training is limited to data captured from a particular 3-D sensor the quantity of training data may be severly limited. In this paper, we show how a crucial a spect of 3-D information–object and feature absolute size–can be added to models learned from commonly available online imagery, without use of any 3-D sensing or re- construction at training time. Such models can be utilized at test time t ogether with explicit 3-D sensing to perform robust search. Our model uses a “2. 1D” local feature, which combines traditional appearance gradient statistics wit h an estimate of average absolute depth within the local window. We show how cat egory size information can be obtained from online images by exploiting relative ly unbiquitous metadata fields specifying camera intrinstics. We develop an effi cient metric branch-and-bound algorithm for our search task, imposing 3-D size c onstraints as part of an optimal search for a set of features which indicate the presence of a category. Experiments on test scenes captured with a traditional stereo rig are shown, exploiting training data from from purely monocular source s with associated EXIF metadata.
****
Near-Optimal Bayesian Active Learning with Noisy Observations
Daniel Golovin, Andreas Krause, Debajyoti Ray
We tackle the fundamental problem of Bayesian active learning with noise, where we need to adaptively select from a number of expensive tests in order to identi fy an unknown hypothesis sampled from a known prior distribution. In the case of noise-free observations, a greedy algorithm called generalized binary search (G BS) is known to perform near-optimally. We show that if the observations are noi sy, perhaps surprisingly, GBS can perform very poorly. We develop EC2, a novel, greedy active learning algorithm and prove that it is competitive with the optim al policy, thus obtaining the first competitiveness guarantees for Bayesian acti ve learning with noisy observations. Our bounds rely on a recently discovered di minishing returns property called adaptive submodularity, generalizing the class ical notion of submodular set functions to adaptive policies. Our results hold e ven if the tests have non–uniform cost and their noise is correlated. We also pr opose EffECXtive, a particularly fast approximation of EC2, and evaluate it on a Bayesian experimental design problem involving human subjects, intended to teas e apart competing economic theories of how people make decisions under uncertain ty.
****
Probabilistic Belief Revision with Structural Constraints
Peter Jones, Venkatesh Saligrama, Sanjoy Mitter
Experts (human or computer) are often required to assess the probability of unce rtain events. When a collection of experts independently assess events that are structurally interrelated, the resulting assessment may violate fundamental laws of probability. Such an assessment is termed incoherent. In this work we invest igate how the problem of incoherence may be affected by allowing experts to spec ify likelihood models and then update their assessments based on the realization of a globally-observable random sequence.
****
Structured Determinantal Point Processes
Alex Kulesza, Ben Taskar
We present a novel probabilistic model for distributions over sets of structures -- for example, sets of sequences, trees, or graphs. The critical characteristi c of our model is a preference for diversity: sets containing dissimilar structu res are more likely. Our model is a marriage of structured probabilistic models, like Markov random fields and context free grammars, with determinantal point p rocesses, which arise in quantum physics as models of particles with repulsive i nteractions. We extend the determinantal point process model to handle an expone ntially-sized set of particles (structures) via a natural factorization of the m

odel into parts. We show how this factorization leads to tractable algorithms fo r exact inference, including computing marginals, computing conditional probabil ities, and sampling. Our algorithms exploit a novel polynomially-sized dual repr esentation of determinantal point processes, and use message passing over a spec ial semiring to compute relevant quantities. We illustrate the advantages of the model on tracking and articulated pose estimation problems.
****

b-Bit Minwise Hashing for Estimating Three-Way Similarities
Ping Li, Arnd Konig, Wenhao Gui
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Feature Transitions with Saccadic Search: Size, Color, and Orientation Are Not A like
Stella Yu
Size, color, and orientation have long been considered elementary features whose attributes are extracted in parallel and available to guide the deployment of a ttention. If each is processed in the same fashion with simply a different set o f local detectors, one would expect similar search behaviours on localizing an e quivalent flickering change among identically laid out disks. We analyze feature transitions associated with saccadic search and find out that size, color, and orientation are not alike in dynamic attribute processing over time. The Markovi an feature transition is attractive for size, repulsive for color, and largely r eversible for orientation.
****

Auto-Regressive HMM Inference with Incomplete Data for Short-Horizon Wind Foreca sting
Chris Barber, Joseph Bockhorst, Paul Roebber
Accurate short-term wind forecasts (STWFs), with time horizons from 0.5 to 6 hou rs, are essential for efficient integration of wind power to the electrical powe r grid. Physical models based on numerical weather predictions are currently not competitive, and research on machine learning approaches is ongoing. Two major challenges confronting these efforts are missing observations and weather-regime induced dependency shifts among wind variables at geographically distributed si tes. In this paper we introduce approaches that address both of these challenges . We describe a new regime-aware approach to STWF that use auto-regressive hidde n Markov models (AR-HMM), a subclass of conditional linear Gaussian (CLG) models . Although AR-HMMs are a natural representation for weather regimes, as with CLG models in general, exact inference is NP-hard when observations are missing (Le rner and Parr, 2001). Because of this high cost, we introduce a simple approxima te inference method for AR-HMMs, which we believe has applications to other sequ ential and temporal problem domains that involve continuous variables. In an emp irical evaluation on publicly available wind data from two geographically distin ct regions, our approach makes significantly more accurate predictions than base line models, and uncovers meteorologically relevant regimes.
****

Link Discovery using Graph Feature Tracking
Emile Richard, Nicolas Baskiotis, Theodoros Evgeniou, Nicolas Vayatis
We consider the problem of discovering links of an evolving undirected graph giv en a series of past snapshots of that graph. The graph is observed through the t ime sequence of its adjacency matrix and only the presence of edges is observed. The absence of an edge on a certain snapshot cannot be distinguished from a mis sing entry in the adjacency matrix. Additional information can be provided by ex amining the dynamics of the graph through a set of topological features, such as the degrees of the vertices. We develop a novel methodology by building on bot h static matrix completion methods and the estimation of the future state of rel evant graph features. Our procedure relies on the formulation of an optimization problem which can be approximately solved by a fast alternating linearized algo

rithm whose properties are examined. We show experiments with both simulated an
d real data which reveal the interest of our methodology.
****
A VLSI Implementation of the Adaptive Exponential Integrate-and-Fire Neuron Mode
l
Sebastian Millner, Andreas Grübl, Karlheinz Meier, Johannes Schemmel, Marc-olivi
er Schwartz
We describe an accelerated hardware neuron being capable of emulating the adap-t
ive exponential integrate-and-fire neuron model. Firing patterns of the membrane
 stimulated by a step current are analyzed in transistor level simulation and in
 silicon on a prototype chip. The neuron is destined to be the hardware neuron o
f a highly integrated wafer-scale system reaching out for new computational para
digms and opening new experimentation possibilities. As the neuron is dedicated
as a universal device for neuroscientific experiments, the focus lays on paramet
erizability and reproduction of the analytical model.
****
Sparse Inverse Covariance Selection via Alternating Linearization Methods
Katya Scheinberg, Shiqian Ma, Donald Goldfarb
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Constructing Skill Trees for Reinforcement Learning Agents from Demonstration Tr
ajectories
George Konidaris, Scott Kuindersma, Roderic Grupen, Andrew Barto
We introduce CST, an algorithm for constructing skill trees from demonstration t
rajectories in continuous reinforcement learning domains. CST uses a changepoint
 detection method to segment each trajectory into a skill chain by detecting a c
hange of appropriate abstraction, or that a segment is too complex to model as a
 single skill. The skill chains from each trajectory are then merged to form a s
kill tree. We demonstrate that CST constructs an appropriate skill tree that can
 be further refined through learning in a challenging continuous domain, and tha
t it can be used to segment demonstration trajectories on a mobile manipulator i
nto chains of skills where each skill is assigned an appropriate abstraction.
****
Trading off Mistakes and Don't-Know Predictions
Amin Sayedi, Morteza Zadimoghaddam, Avrim Blum
We discuss an online learning framework in which the agent is allowed to say I d
on't know'' as well as making incorrect predictions on given examples. We analyz
e the trade off between sayingI don't know'' and making mistakes. If the number
of don't know predictions is forced to be zero, the model reduces to the well-kn
own mistake-bound model introduced by Littlestone [Lit88]. On the other hand, if
 no mistakes are allowed, the model reduces to KWIK framework introduced by Li e
t. al. [LLW08]. We propose a general, though inefficient, algorithm for general
finite concept classes that minimizes the number of don't-know predictions if a
certain number of mistakes are allowed. We then present specific polynomial-time
 algorithms for the concept classes of monotone disjunctions and linear separato
rs.
****
Evaluation of Rarity of Fingerprints in Forensics
Chang Su, Sargur Srihari
A method for computing the rarity of latent fingerprints represented by minutiae
 is given. It allows determining the probability of finding a match for an evide
nce print in a database of n known prints. The probability of random corresponde
nce between evidence and database is determined in three procedural steps. In th
e registration step the latent print is aligned by finding its core point; which
 is done using a procedure based on a machine learning approach based on Gaussia
n processes. In the evidence probability evaluation step a generative model base
d on Bayesian networks is used to determine the probability of the evidence; it

takes into account both the dependency of each minutia on nearby minutiae and th
e confidence of their presence in the evidence. In the specific probability of r
andom correspondence step the evidence probability is used to determine the prob
ability of match among n for a given tolerance; the last evaluation is similar t
o the birthday correspondence probability for a specific birthday. The generativ
e model is validated using a goodness-of-fit test evaluated with a standard data
base of fingerprints. The probability of random correspondence for several laten
t fingerprints are evaluated for varying numbers of minutiae.
****
(RF)^2 -- Random Forest Random Field
Nadia Payet, Sinisa Todorovic
We combine random forest (RF) and conditional random field (CRF) into a new comp
utational framework, called random forest random field (RF)^2. Inference of (RF)
^2 uses the Swendsen-Wang cut algorithm, characterized by Metropolis-Hastings ju
mps. A jump from one state to another depends on the ratio of the proposal distr
ibutions, and on the ratio of the posterior distributions of the two states. Pri
or work typically resorts to a parametric estimation of these four distributions
, and then computes their ratio. Our key idea is to instead directly estimate th
ese ratios using RF. RF collects in leaf nodes of each decision tree the class h
istograms of training examples. We use these class histograms for a non-parametr
ic estimation of the distribution ratios. We derive the theoretical error bounds
 of a two-class (RF)^2. (RF)^2 is applied to a challenging task of multiclass ob
ject recognition and segmentation over a random field of input image regions. In
 our empirical evaluation, we use only the visual information provided by image
regions (e.g., color, texture, spatial layout), whereas the competing methods ad
ditionally use higher-level cues about the horizon location and 3D layout of sur
faces in the scene. Nevertheless, (RF)^2 outperforms the state of the art on ben
chmark datasets, in terms of accuracy and computation time.
****
Online Learning in The Manifold of Low-Rank Matrices
Uri Shalit, Daphna Weinshall, Gal Chechik
When learning models that are represented in matrix forms, enforcing a low-ran
k constraint can dramatically improve the memory and run time complexity, whil
e providing a natural regularization of the model. However, naive approaches
for minimizing functions over the set of low-rank matrices are either prohibit
ively time consuming (repeated singular value decomposition of the matrix) or
  numerically unstable (optimizing a factored representation of the low rank m
atrix). We build on recent advances in optimization over manifolds, and descri
be an iterative online learning procedure, consisting of a gradient step, foll
owed by a second-order retraction back to the manifold. While the ideal retrac
tion is hard to compute, and so is the projection operator that approximates i
t, we describe another second-order retraction that can be computed efficientl
y, with run time and memory complexity of O((n+m)k) for a rank-k matrix of d
imension m x n, given rank one gradients. We use this algorithm, LORETA, to l
earn a matrix-form similarity measure over pairs of documents represented as h
igh dimensional vectors. LORETA improves the mean average precision over a pas
sive- aggressive approach in a factorized model, and also improves over a fu
ll model trained over pre-selected features using the same memory requirements
. LORETA also showed consistent improvement over standard methods in a large (
1600 classes) multi-label image classification task.
****
Variational bounds for mixed-data factor analysis
Mohammad Emtiyaz E. Khan, Guillaume Bouchard, Kevin P. Murphy, Benjamin M. Marli
n
We propose a new variational EM algorithm for fitting factor analysis models wit
h mixed continuous and categorical observations. The algorithm is based on a sim
ple quadratic bound to the log-sum-exp function. In the special case of fully o
bserved binary data, the bound we propose is significantly faster than previous
variational methods. We show that EM is significantly more robust in the presenc
e of missing data compared to treating the latent factors as parameters, which i

s the approach used by exponential family PCA and other related matrix-factoriza tion methods. A further benefit of the variational approach is that it can easi ly be extended to the case of mixtures of factor analyzers, as we show. We prese nt results on synthetic and real data sets demonstrating several desirable prope rties of our proposed method.
****

Copula Bayesian Networks
Gal Elidan
We present the Copula Bayesian Network model for representing multivariate conti nuous distributions. Our approach builds on a novel copula-based parameterizatio n of a conditional density that, joined with a graph that encodes independencies , offers great flexibility in modeling high-dimensional densities, while maintai ning control over the form of the univariate marginals. We demonstrate the advan tage of our framework for generalization over standard Bayesian networks as well as tree structured copula models for varied real-life domains that are of subst antially higher dimension than those typically considered in the copula literatu re.
****

Evidence-Specific Structures for Rich Tractable CRFs
Anton Chechetka, Carlos Guestrin
We present a simple and effective approach to learning tractable conditional ran dom fields with structure that depends on the evidence. Our approach retains the advantages of tractable discriminative models, namely efficient exact inference and exact parameter learning. At the same time, our algorithm does not suffer a large expressive power penalty inherent to fixed tractable structures. On real- life relational datasets, our approach matches or exceeds state of the art accur acy of the dense models, and at the same time provides an order of magnitude spe edup
****

Beyond Actions: Discriminative Models for Contextual Group Activities
Tian Lan, Yang Wang, Weilong Yang, Greg Mori
We propose a discriminative model for recognizing group activities. Our model jo intly captures the group activity, the individual person actions, and the intera ctions among them. Two new types of contextual information, group-person interac tion and person-person interaction, are explored in a latent variable framework. Different from most of the previous latent structured models which assume a pre defined structure for the hidden layer, e.g. a tree structure, we treat the stru cture of the hidden layer as a latent variable and implicitly infer it during le arning and inference. Our experimental results demonstrate that by inferring thi s contextual information together with adaptive structures, the proposed model c an significantly improve activity recognition performance.
****

Decoding Ipsilateral Finger Movements from ECoG Signals in Humans
Yuzong Liu, Mohit Sharma, Charles Gaona, Jonathan Breshears, Jarod Roland, Zacha ry Freudenburg, Eric Leuthardt, Kilian Q. Weinberger
Several motor related Brain Computer Interfaces (BCIs) have been developed over the years that use activity decoded from the contralateral hemisphere to operate devices. Many recent studies have also talked about the importance of ipsilater al activity in planning of motor movements. For successful upper limb BCIs, it i s important to decode finger movements from brain activity. This study uses ipsi lateral cortical signals from humans (using ECoG) to decode finger movements. We demonstrate, for the first time, successful finger movement detection using mac hine learning algorithms. Our results show high decoding accuracies in all cases which are always above chance. We also show that significant accuracies can be achieved with the use of only a fraction of all the features recorded and that t hese core features also make sense physiologically. The results of this study ha ve a great potential in the emerging world of motor neuroprosthetics and other B CIs.
****

New Adaptive Algorithms for Online Classification

Francesco Orabona, Koby Crammer
We propose a general framework to online learning for classification problems with time-varying potential functions in the adversarial setting. This framework allows to design and prove relative mistake bounds for any generic loss function. The mistake bounds can be specialized for the hinge loss, allowing to recover and improve the bounds of known online classification algorithms. By optimizing the general bound we derive a new online classification algorithm, called NAROW, that hybridly uses adaptive- and fixed- second order information. We analyze the properties of the algorithm and illustrate its performance using synthetic dataset.
****

Phoneme Recognition with Large Hierarchical Reservoirs
Fabian Triefenbach, Azarakhsh Jalalvand, Benjamin Schrauwen, Jean-pierre Martens
Automatic speech recognition has gradually improved over the years, but the reliable recognition of unconstrained speech is still not within reach. In order to achieve a breakthrough, many research groups are now investigating new methodologies that have potential to outperform the Hidden Markov Model technology that is at the core of all present commercial systems. In this paper, it is shown that the recently introduced concept of Reservoir Computing might form the basis of such a methodology. In a limited amount of time, a reservoir system that can recognize the elementary sounds of continuous speech has been built. The system already achieves a state-of-the-art performance, and there is evidence that the margin for further improvements is still significant.
****

Learning Multiple Tasks using Manifold Regularization
Arvind Agarwal, Samuel Gerber, Hal Daume
We present a novel method for multitask learning (MTL) based on {\it manifold regularization}: assume that all task parameters lie on a manifold. This is the generalization of a common assumption made in the existing literature: task parameters share a common {\it linear} subspace. One proposed method uses the projection distance from the manifold to regularize the task parameters. The manifold structure and the task parameters are learned using an alternating optimization framework. When the manifold structure is fixed, our method decomposes across tasks which can be learnt independently. An approximation of the manifold regularization scheme is presented that preserves the convexity of the single task learning problem, and makes the proposed MTL framework efficient and easy to implement. We show the efficacy of our method on several datasets.
****

Group Sparse Coding with a Laplacian Scale Mixture Prior
Pierre Garrigues, Bruno Olshausen
We propose a class of sparse coding models that utilizes a Laplacian Scale Mixture (LSM) prior to model dependencies among coefficients. Each coefficient is modeled as a Laplacian distribution with a variable scale parameter, with a Gamma distribution prior over the scale parameter. We show that, due to the conjugacy of the Gamma prior, it is possible to derive efficient inference procedures for both the coefficients and the scale parameter. When the scale parameters of a group of coefficients are combined into a single variable, it is possible to describe the dependencies that occur due to common amplitude fluctuations among coefficients, which have been shown to constitute a large fraction of the redundancy in natural images. We show that, as a consequence of this group sparse coding, the resulting inference of the coefficients follows a divisive normalization rule, and that this may be efficiently implemented a network architecture similar to that which has been proposed to occur in primary visual cortex. We also demonstrate improvements in image coding and compressive sensing recovery using the LSM model.
****

Fractionally Predictive Spiking Neurons
Jaldert Rombouts, Sander Bohte
Recent experimental work has suggested that the neural firing rate can be interpreted as a fractional derivative, at least when signal variation induces neural

adaptation. Here, we show that the actual neural spike-train itself can be consi
dered as the fractional derivative, provided that the neural signal is approxima
ted by a sum of power-law kernels. A simple standard thresholding spiking neuron
 suffices to carry out such an approximation, given a suitable refractory respon
se. Empirically, we find that the online approximation of signals with a sum of
power-law kernels is beneficial for encoding signals with slowly varying compone
nts, like long-memory self-similar signals. For such signals, the online power-l
aw kernel approximation typically required less than half the number of spikes f
or similar SNR as compared to sums of similar but exponentially decaying kernels
. As power-law kernels can be accurately approximated using sums or cascades of
weighted exponentials, we demonstrate that the corresponding decoding of spike-t
rains by a receiving neuron allows for natural and transparent temporal signal f
iltering by tuning the weights of the decoding kernel.
****

Stability Approach to Regularization Selection (StARS) for High Dimensional Grap
hical Models
Han Liu, Kathryn Roeder, Larry Wasserman
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

More data means less inference: A pseudo-max approach to structured learning
David Sontag, Ofer Meshi, Amir Globerson, Tommi Jaakkola
The problem of learning to predict structured labels is of key importance in man
y applications. However, for general graph structure both learning and inference
 in this setting are intractable. Here we show that it is possible to circumvent
 this difficulty when the input distribution is rich enough via a method similar
 in spirit to pseudo-likelihood. We show how our new method achieves consistency
, and illustrate empirically that it indeed performs as well as exact methods wh
en sufficiently large training sets are used.
****

Lifted Inference Seen from the Other Side : The Tractable Features
Abhay Jha, Vibhav Gogate, Alexandra Meliou, Dan Suciu
Lifted inference algorithms for representations that combine first-order logic a
nd probabilistic graphical models have been the focus of much recent research. A
ll lifted algorithms developed to date are based on the same underlying idea: ta
ke a standard probabilistic inference algorithm (e.g., variable elimination, bel
ief propagation etc.) and improve its efficiency by exploiting repeated structur
e in the first-order model. In this paper, we propose an approach from the other
 side in that we use techniques from logic for probabilistic inference. In parti
cular, we define a set of rules that look only at the logical representation to
identify models for which exact efficient inference is possible. We show that ou
r rules yield several new tractable classes that cannot be solved efficiently by
 any of the existing techniques.
****

Predictive State Temporal Difference Learning
Byron Boots, Geoffrey J. Gordon
We propose a new approach to value function approximation which combines linear
temporal difference reinforcement learning with subspace identification. In prac
tical applications, reinforcement learning (RL) is complicated by the fact that
state is either high-dimensional or partially observable. Therefore, RL methods
are designed to work with features of state rather than state itself, and the su
ccess or failure of learning is often determined by the suitability of the selec
ted features. By comparison, subspace identification (SSID) methods are designed
 to select a feature set which preserves as much information as possible about s
tate. In this paper we connect the two approaches, looking at the problem of rei
nforcement learning with a large set of features, each of which may only be marg
inally useful for value function approximation. We introduce a new algorithm for
 this situation, called Predictive State Temporal Difference (PSTD) learning. As

in SSID for predictive state representations, PSTD finds a linear compression operator that projects a large set of features down to a small set that preserves the maximum amount of predictive information. As in RL, PSTD then uses a Bellman recursion to estimate a value function. We discuss the connection between PSTD and prior approaches in RL and SSID. We prove that PSTD is statistically consistent, perform several experiments that illustrate its properties, and demonstrate its potential on a difficult optimal stopping problem.
****

Identifying Dendritic Processing
Aurel A. Lazar, Yevgeniy Slutskiy
In system identification both the input and the output of a system are available to an observer and an algorithm is sought to identify parameters of a hypothesized model of that system. Here we present a novel formal methodology for identifying dendritic processing in a neural circuit consisting of a linear dendritic processing filter in cascade with a spiking neuron model. The input to the circuit is an analog signal that belongs to the space of bandlimited functions. The output is a time sequence associated with the spike train. We derive an algorithm for identification of the dendritic processing filter and reconstruct its kernel with arbitrary precision.
****

On a Connection between Importance Sampling and the Likelihood Ratio Policy Gradient
Tang Jie, Pieter Abbeel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Functional Geometry Alignment and Localization of Brain Areas
Georg Langs, Yanmei Tie, Laura Rigolo, Alexandra Golby, Polina Golland
Matching functional brain regions across individuals is a challenging task, largely due to the variability in their location and extent. It is particularly difficult, but highly relevant, for patients with pathologies such as brain tumors, which can cause substantial reorganization of functional systems. In such cases spatial registration based on anatomical data is only of limited value if the goal is to establish correspondences of functional areas among different individuals, or to localize potentially displaced active regions. Rather than rely on spatial alignment, we propose to perform registration in an alternative space whose geometry is governed by the functional interaction patterns in the brain. We first embed each brain into a functional map that reflects connectivity patterns during a fMRI experiment. The resulting functional maps are then registered, and the obtained correspondences are propagated back to the two brains. In application to a language fMRI experiment, our preliminary results suggest that the proposed method yields improved functional correspondences across subjects. This advantage is pronounced for subjects with tumors that affect the language areas and thus cause spatial reorganization of the functional regions.
****

Multi-View Active Learning in the Non-Realizable Case
Wei Wang, Zhi-Hua Zhou
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Epitome driven 3-D Diffusion Tensor image segmentation: on extracting specific structures
Kamiya Motwani, Nagesh Adluru, Chris Hinrichs, Andrew Alexander, Vikas Singh
We study the problem of segmenting specific white matter structures of interest from Diffusion Tensor (DT-MR) images of the human brain. This is an important requirement in many Neuroimaging studies: for instance, to evaluate whether a brai

n structure exhibits group level differences as a function of disease in a set o
f images. Typically, interactive expert guided segmentation has been the method
of choice for such applications, but this is tedious for large datasets common t
oday. To address this problem, we endow an image segmentation algorithm with 'ad
vice' encoding some global characteristics of the region(s) we want to extract.
This is accomplished by constructing (using expert-segmented images) an epitome
of a specific region - as a histogram over a bag of 'words' (e.g.,suitable featu
re descriptors). Now, given such a representation, the problem reduces to segmen
ting new brain image with additional constraints that enforce consistency betwee
n the segmented foreground and the pre-specified histogram over features. We pre
sent combinatorial approximation algorithms to incorporate such domain specific
constraints for Markov Random Field (MRF) segmentation. Making use of recent res
ults on image co-segmentation, we derive effective solution strategies for our p
roblem. We provide an analysis of solution quality, and present promising experi
mental evidence showing that many structures of interest in Neuroscience can be
extracted reliably from 3-D brain image volumes using our algorithm.
****

Over-complete representations on recurrent neural networks can support persisten
t percepts

Shaul Druckmann, Dmitri Chklovskii

A striking aspect of cortical neural networks is the divergence of a relatively
small number of input channels from the peripheral sensory apparatus into a larg
e number of cortical neurons, an over-complete representation strategy. Cortical
neurons are then connected by a sparse network of lateral synapses. Here we pro
pose that such architecture may increase the persistence of the representation o
f an incoming stimulus, or a percept. We demonstrate that for a family of networ
ks in which the receptive field of each neuron is re-expressed by its outgoing c
onnections, a represented percept can remain constant despite changing activity.
We term this choice of connectivity REceptive FIeld REcombination (REFIRE) netw
orks. The sparse REFIRE network may serve as a high-dimensional integrator and a
biologically plausible model of the local cortical circuit.
****

Non-Stochastic Bandit Slate Problems

Satyen Kale, Lev Reyzin, Robert E. Schapire

We consider bandit problems, motivated by applications in online advertising and
news story selection, in which the learner must repeatedly select a slate, that
is, a subset of size s from K possible actions, and then receives rewards for j
ust the selected actions. The goal is to minimize the regret with respect to tot
al reward of the best slate computed in hindsight. We consider unordered and ord
ered versions of the problem, and give efficient algorithms which have regret O(
sqrt(T)), where the constant depends on the specific nature of the problem. We a
lso consider versions of the problem where we have access to a number of policie
s which make recommendations for slates in every round, and give algorithms with
O(sqrt(T)) regret for competing with the best such policy as well. We make use
of the technique of relative entropy projections combined with the usual multipl
icative weight update algorithm to obtain our algorithms.
****

Large Margin Learning of Upstream Scene Understanding Models

Jun Zhu, Li-jia Li, Li Fei-fei, Eric Xing

Upstream supervised topic models have been widely used for complicated scene und
erstanding. However, existing maximum likelihood estimation (MLE) schemes can ma
ke the prediction model learning independent of latent topic discovery and resul
t in an imbalanced prediction rule for scene classification. This paper presents
a joint max-margin and max-likelihood learning method for upstream scene unders
tanding models, in which latent topic discovery and prediction model estimation
are closely coupled and well-balanced. The optimization problem is efficiently s
olved with a variational EM procedure, which iteratively solves an online loss-a
ugmented SVM. We demonstrate the advantages of the large-margin approach on both
an 8-category sports dataset and the 67-class MIT indoor scene dataset for scen
e categorization.

****
Switched Latent Force Models for Movement Segmentation
Mauricio Alvarez, Jan Peters, Neil Lawrence, Bernhard Schölkopf
Latent force models encode the interaction between multiple related dynamical sy
stems in the form of a kernel or covariance function. Each variable to be modele
d is represented as the output of a differential equation and each differential
equation is driven by a weighted sum of latent functions with uncertainty given
by a Gaussian process prior. In this paper we consider employing the latent forc
e model framework for the problem of determining robot motor primitives. To deal
with discontinuities in the dynamical systems or the latent driving force we in
troduce an extension of the basic latent force model, that switches between diff
erent latent functions and potentially different dynamical systems. This creates
a versatile representation for robot movements that can capture discrete change
s and non-linearities in the dynamics. We give illustrative examples on both syn
thetic data and for striking movements recorded using a Barrett WAM robot as hap
tic input device. Our inspiration is robot motor primitives, but we expect our m
odel to have wide application for dynamical systems including models for human m
otion capture data and systems biology.
****
Adaptive Multi-Task Lasso: with Application to eQTL Detection
Seunghak Lee, Jun Zhu, Eric Xing
To understand the relationship between genomic variations among population and c
omplex diseases, it is essential to detect eQTLs which are associated with pheno
typic effects. However, detecting eQTLs remains a challenge due to complex under
lying mechanisms and the very large number of genetic loci involved compared to
the number of samples. Thus, to address the problem, it is desirable to take adv
antage of the structure of the data and prior information about genomic location
s such as conservation scores and transcription factor binding sites. In this p
aper, we propose a novel regularized regression approach for detecting eQTLs whi
ch takes into account related traits simultaneously while incorporating many reg
ulatory features. We first present a Bayesian network for a multi-task learning
problem that includes priors on SNPs, making it possible to estimate the signifi
cance of each covariate adaptively. Then we find the maximum a posteriori (MAP)
estimation of regression coefficients and estimate weights of covariates jointly
. This optimization procedure is efficient since it can be achieved by using con
vex optimization and a coordinate descent procedure iteratively. Experimental re
sults on simulated and real yeast datasets confirm that our model outperforms pr
evious methods for finding eQTLs.
****
Categories and Functional Units: An Infinite Hierarchical Model for Brain Activa
tions
Danial Lashkari, Ramesh Sridharan, Polina Golland
We present a model that describes the structure in the responses of different br
ain areas to a set of stimuli in terms of stimulus categories" (clusters of stim
uli) and "functional units" (clusters of voxels). We assume that voxels within a
unit respond similarly to all stimuli from the same category, and design a nonp
arametric hierarchical model to capture inter-subject variability among the unit
s. The model explicitly captures the relationship between brain activations and
fMRI time courses. A variational inference algorithm derived based on the model
can learn categories, units, and a set of unit-category activation probabilities
from data. When applied to data from an fMRI study of object recognition, the m
ethod finds meaningful and consistent clusterings of stimuli into categories and
voxels into units."
****
Random Projection Trees Revisited
Aman Dhesi, Purushottam Kar
The Random Projection Tree (RPTree) structures proposed in [Dasgupta-Freund-STOC
-08] are space partitioning data structures that automatically adapt to various
notions of intrinsic dimensionality of data. We prove new results for both the R
PTree-Max and the RPTree-Mean data structures. Our result for RPTree-Max gives a

near-optimal bound on the number of levels required by this data structure to r
educe the size of its cells by a factor s >= 2. We also prove a packing lemma fo
r this data structure. Our final result shows that low-dimensional manifolds pos
sess bounded Local Covariance Dimension. As a consequence we show that RPTree-Me
an adapts to manifold dimension as well.
****

Joint Analysis of Time-Evolving Binary Matrices and Associated Documents
Eric Wang, Dehong Liu, Jorge Silva, Lawrence Carin, David Dunson
We consider problems for which one has incomplete binary matrices that evolve wi
th time (e.g., the votes of legislators on particular legislation, with each yea
r characterized by a different such matrix). An objective of such analysis is to
infer structure and inter-relationships underlying the matrices, here defined b
y latent features associated with each axis of the matrix. In addition, it is as
sumed that documents are available for the entities associated with at least one
of the matrix axes. By jointly analyzing the matrices and documents, one may be
used to inform the other within the analysis, and the model offers the opportun
ity to predict matrix values (e.g., votes) based only on an associated document
(e.g., legislation). The research presented here merges two areas of machine-lea
rning that have previously been investigated separately: incomplete-matrix analy
sis and topic modeling. The analysis is performed from a Bayesian perspective, w
ith efficient inference constituted via Gibbs sampling. The framework is demonst
rated by considering all voting data and available documents (legislation) durin
g the 220-year lifetime of the United States Senate and House of Representatives
.
****

Discriminative Clustering by Regularized Information Maximization
Andreas Krause, Pietro Perona, Ryan Gomes
Is there a principled way to learn a probabilistic discriminative classifier fro
m an unlabeled data set? We present a framework that simultaneously clusters the
data and trains a discriminative classifier. We call it Regularized Information
Maximization (RIM). RIM optimizes an intuitive information-theoretic objective
function which balances class separation, class balance and classifier complexit
y. The approach can flexibly incorporate different likelihood functions, express
prior assumptions about the relative size of different classes and incorporate
partial labels for semi-supervised learning. In particular, we instantiate the f
ramework to unsupervised, multi-class kernelized logistic regression. Our empiri
cal evaluation indicates that RIM outperforms existing methods on several real d
ata sets, and demonstrates that RIM is an effective model selection method.
****

Learning to localise sounds with spiking neural networks
Dan Goodman, Romain Brette
To localise the source of a sound, we use location-specific properties of the si
gnals received at the two ears caused by the asymmetric filtering of the origina
l sound by our head and pinnae, the head-related transfer functions (HRTFs). The
se HRTFs change throughout an organism's lifetime, during development for exampl
e, and so the required neural circuitry cannot be entirely hardwired. Since HRTF
s are not directly accessible from perceptual experience, they can only be infer
red from filtered sounds. We present a spiking neural network model of sound loc
alisation based on extracting location-specific synchrony patterns, and a simple
supervised algorithm to learn the mapping between synchrony patterns and locati
ons from a set of example sounds, with no previous knowledge of HRTFs. After lea
rning, our model was able to accurately localise new sounds in both azimuth and
elevation, including the difficult task of distinguishing sounds coming from the
front and back.
****

Dynamic Infinite Relational Model for Time-varying Relational Data Analysis
Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, Joshua Tenenbaum
We propose a new probabilistic model for analyzing dynamic evolutions of relatio
nal data, such as additions, deletions and split & merge, of relation clusters l
ike communities in social networks. Our proposed model abstracts observed time-v

arying object-object relationships into relationships between object clusters. We extend the infinite Hidden Markov model to follow dynamic and time-sensitive changes in the structure of the relational data and to estimate a number of clusters simultaneously. We show the usefulness of the model through experiments with synthetic and real-world data sets.
****

Exact learning curves for Gaussian process regression on large random graphs
Matthew Urry, Peter Sollich
We study learning curves for Gaussian process regression which characterise performance in terms of the Bayes error averaged over datasets of a given size. Whilst learning curves are in general very difficult to calculate we show that for discrete input domains, where similarity between input points is characterised in terms of a graph, accurate predictions can be obtained. These should in fact become exact for large graphs drawn from a broad range of random graph ensembles with arbitrary degree distributions where each input (node) is connected only to a finite number of others. The method is based on translating the appropriate belief propagation equations to the graph ensemble. We demonstrate the accuracy of the predictions for Poisson (Erdos-Renyi) and regular random graphs, and discuss when and why previous approximations to the learning curve fail.
****

Sparse Instrumental Variables (SPIV) for Genome-Wide Studies
Paul Mckeigue, Jon Krohn, Amos J. Storkey, Felix Agakov
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Natural Policy Gradient Methods with Parameter-based Exploration for Control Tasks
Atsushi Miyamae, Yuichi Nagata, Isao Ono, Shigenobu Kobayashi
In this paper, we propose an efficient algorithm for estimating the natural policy gradient with parameter-based exploration; this algorithm samples directly in the parameter space. Unlike previous methods based on natural gradients, our algorithm calculates the natural policy gradient using the inverse of the exact Fisher information matrix. The computational cost of this algorithm is equal to that of conventional policy gradients whereas previous natural policy gradient methods have a prohibitive computational cost. Experimental results show that the proposed method outperforms several policy gradient methods.
****

Kernel Descriptors for Visual Recognition
Liefeng Bo, Xiaofeng Ren, Dieter Fox
The design of low-level image features is critical for computer vision algorithms. Orientation histograms, such as those in SIFT~\cite{Lowe2004Distinctive} and HOG~\cite{Dalal2005Histograms}, are the most successful and popular features for visual object and scene recognition. We highlight the kernel view of orientation histograms, and show that they are equivalent to a certain type of match kernels over image patches. This novel view allows us to design a family of kernel descriptors which provide a unified and principled framework to turn pixel attributes (gradient, color, local binary pattern, \etc) into compact patch-level features. In particular, we introduce three types of match kernels to measure similarities between image patches, and construct compact low-dimensional kernel descriptors from these match kernels using kernel principal component analysis (KPCA)~\cite{Scholkopf1998Nonlinear}. Kernel descriptors are easy to design and can turn any type of pixel attribute into patch-level features. They outperform carefully tuned and sophisticated features including SIFT and deep belief networks. We report superior performance on standard image classification benchmarks: Scene-15, Caltech-101, CIFAR10 and CIFAR10-ImageNet.
****

Computing Marginal Distributions over Continuous Markov Networks for Statistical Relational Learning

Matthias Broecheler, Lise Getoor
Continuous Markov random fields are a general formalism to model joint probabili ty distributions over events with continuous outcomes. We prove that marginal co mputation for constrained continuous MRFs is #P-hard in general and present a po lynomial-time approximation scheme under mild assumptions on the structure of th e random field. Moreover, we introduce a sampling algorithm to compute marginal distributions and develop novel techniques to increase its efficiency. Continuou s MRFs are a general purpose probabilistic modeling tool and we demonstrate how they can be applied to statistical relational learning. On the problem of collec tive classification, we evaluate our algorithm and show that the standard deviat ion of marginals serves as a useful measure of confidence.
****

Gaussian Process Preference Elicitation
Shengbo Guo, Scott Sanner, Edwin V. Bonilla
Bayesian approaches to preference elicitation (PE) are particularly attractive d ue to their ability to explicitly model uncertainty in users' latent utility fun ctions. However, previous approaches to Bayesian PE have ignored the important p roblem of generalizing from previous users to an unseen user in order to reduce the elicitation burden on new users. In this paper, we address this deficiency b y introducing a Gaussian Process (GP) prior over users' latent utility functions on the joint space of user and item features. We learn the hyper-parameters of this GP on a set of preferences of previous users and use it to aid in the elici tation process for a new user. This approach provides a flexible model of a mult i-user utility function, facilitates an efficient value of information (VOI) heu ristic query selection strategy, and provides a principled way to incorporate th e elicitations of multiple users back into the model. We show the effectiveness of our method in comparison to previous work on a real dataset of user preferenc es over sushi types.
****

A Theory of Multiclass Boosting
Indraneel Mukherjee, Robert E. Schapire
Boosting combines weak classifiers to form highly accurate predictors. Although the case of binary classification is well understood, in the multiclass setting, the "correct" requirements on the weak classifier, or the notion of the most ef ficient boosting algorithms are missing. In this paper, we create a broad and ge neral framework, within which we make precise and identify the optimal requireme nts on the weak-classifier, as well as design the most effective, in a certain s ense, boosting algorithms that assume such requirements.
****

Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Activ e Learning
Prateek Jain, Sudheendra Vijayanarasimhan, Kristen Grauman
We consider the problem of retrieving the database points nearest to a given {\e m hyperplane} query without exhaustively scanning the database. We propose two h ashing-based solutions. Our first approach maps the data to two-bit binary keys that are locality-sensitive for the angle between the hyperplane normal and a da tabase point. Our second approach embeds the data into a vector space where the Euclidean norm reflects the desired distance between the original points and hyp erplane query. Both use hashing to retrieve near points in sub-linear time. Our first method's preprocessing stage is more efficient, while the second has stron ger accuracy guarantees. We apply both to pool-based active learning: taking the current hyperplane classifier as a query, our algorithm identifies those points (approximately) satisfying the well-known minimal distance-to-hyperplane select ion criterion. We empirically demonstrate our methods' tradeoffs, and show that they make it practical to perform active selection with millions of unlabeled po ints.
****

Bootstrapping Apprenticeship Learning
Abdeslam Boularias, Brahim Chaib-draa
We consider the problem of apprenticeship learning where the examples, demonstra

ted by an expert, cover only a small part of a large state space. Inverse Reinfo rcement Learning (IRL) provides an efficient tool for generalizing the demonstra tion, based on the assumption that the expert is maximizing a utility function t hat is a linear combination of state-action features. Most IRL algorithms use a simple Monte Carlo estimation to approximate the expected feature counts under t he expert's policy. In this paper, we show that the quality of the learned polic ies is highly sensitive to the error in estimating the feature counts. To reduce this error, we introduce a novel approach for bootstrapping the demonstration b y assuming that: (i), the expert is (near-)optimal, and (ii), the dynamics of th e system is known. Empirical results on gridworlds and car racing problems show that our approach is able to learn good policies from a small number of demonstr ations.

****

Co-regularization Based Semi-supervised Domain Adaptation

Abhishek Kumar, Avishek Saha, Hal Daume

This paper presents a co-regularization based approach to semi-supervised domain adaptation. Our proposed approach (EA++) builds on the notion of augmented spac e (introduced in EASYADAPT (EA) [1]) and harnesses unlabeled data in target doma in to further enable the transfer of information from source to target. This sem i-supervised approach to domain adaptation is extremely simple to implement and can be applied as a pre-processing step to any supervised learner. Our theoretic al analysis (in terms of Rademacher complexity) of EA and EA++ show that the hyp othesis class of EA++ has lower complexity (compared to EA) and hence results in tighter generalization bounds. Experimental results on sentiment analysis tasks reinforce our theoretical findings and demonstrate the efficacy of the proposed method when compared to EA as well as a few other baseline approaches.

****

Structured sparsity-inducing norms through submodular functions

Francis Bach

Sparse methods for supervised learning aim at finding good linear predictors fro m as few variables as possible, i.e., with small cardinality of their supports. This combinatorial selection problem is often turned into a convex optimization problem by replacing the cardinality function by its convex envelope (tightest c onvex lower bound), in this case the L1-norm. In this paper, we investigate more general set-functions than the cardinality, that may incorporate prior knowledg e or structural constraints which are common in many applications: namely, we sh ow that for nondecreasing submodular set-functions, the corresponding convex env elope can be obtained from its Lovasz extension, a common tool in submodular ana lysis. This defines a family of polyhedral norms, for which we provide generic a lgorithmic tools (subgradients and proximal operators) and theoretical results (conditions for support recovery or high-dimensional inference). By selecting spe cific submodular functions, we can give a new interpretation to known norms, suc h as those based on rank-statistics or grouped norms with potentially overlappin g groups; we also define new norms, in particular ones that can be used as non-f actorial priors for supervised learning.

****

Optimal Web-Scale Tiering as a Flow Problem

Gilbert Leung, Novi Quadrianto, Kostas Tsioutsiouliklis, Alex Smola

We present a fast online solver for large scale maximum-flow problems as they oc cur in portfolio optimization, inventory management, computer vision, and logist ics. Our algorithm solves an integer linear program in an online fashion. It exp loits total unimodularity of the constraint matrix and a Lagrangian relaxation t o solve the problem as a convex online game. The algorithm generates approximate solutions of max-flow problems by performing stochastic gradient descent on a s et of flows. We apply the algorithm to optimize tier arrangement of over 80 Mill ion web pages on a layered set of caches to serve an incoming query stream optim ally. We provide an empirical demonstration of the effectiveness of our method o n real query-pages data.

****

Slice sampling covariance hyperparameters of latent Gaussian models

Iain Murray, Ryan P. Adams
The Gaussian process (GP) is a popular way to specify dependencies between rando m variables in a probabilistic model. In the Bayesian framework the covariance s tructure can be specified using unknown hyperparameters. Integrating over these hyperparameters considers different possible explanations for the data when maki ng predictions. This integration is often performed using Markov chain Monte Car lo (MCMC) sampling. However, with non-Gaussian observations standard hyperparame ter sampling approaches require careful tuning and may converge slowly. In this paper we present a slice sampling approach that requires little tuning while mix ing well in both strong- and weak-data regimes.
****

Efficient Optimization for Discriminative Latent Class Models
Armand Joulin, Jean Ponce, Francis Bach
Dimensionality reduction is commonly used in the setting of multi-label supervis ed classification to control the learning capacity and to provide a meaningful r epresentation of the data. We introduce a simple forward probabilistic model whi ch is a multinomial extension of reduced rank regression; we show that this mode l provides a probabilistic interpretation of discriminative clustering methods w ith added benefits in terms of number of hyperparameters and optimization. While expectation-maximization (EM) algorithm is commonly used to learn these models, its optimization usually leads to local minimum because it relies on a non-conv ex cost function with many such local minima. To avoid this problem, we introduc e a local approximation of this cost function, which leads to a quadratic non-co nvex optimization problem over a product of simplices. In order to minimize such functions, we propose an efficient algorithm based on convex relaxation and low -rank representation of our data, which allows to deal with large instances. Exp eriments on text document classification show that the new model outperforms oth er supervised dimensionality reduction methods, while simulations on unsupervise d clustering show that our probabilistic formulation has better properties than existing discriminative clustering methods.
****

Universal Kernels on Non-Standard Input Spaces
Andreas Christmann, Ingo Steinwart
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Worst-Case Linear Discriminant Analysis
Yu Zhang, Dit-Yan Yeung
Dimensionality reduction is often needed in many applications due to the high di mensionality of the data involved. In this paper, we first analyze the scatter m easures used in the conventional linear discriminant analysis~(LDA) model and no te that the formulation is based on the average-case view. Based on this analysi s, we then propose a new dimensionality reduction method called worst-case linea r discriminant analysis~(WLDA) by defining new between-class and within-class sc atter measures. This new model adopts the worst-case view which arguably is more suitable for applications such as classification. When the number of training d ata points or the number of features is not very large, we relax the optimizatio n problem involved and formulate it as a metric learning problem. Otherwise, we take a greedy approach by finding one direction of the transformation at a time. Moreover, we also analyze a special case of WLDA to show its relationship with conventional LDA. Experiments conducted on several benchmark datasets demonstrat e the effectiveness of WLDA when compared with some related dimensionality reduc tion methods.
****

Learning Multiple Tasks with a Sparse Matrix-Normal Penalty
Yi Zhang, Jeff Schneider
In this paper, we propose a matrix-variate normal penalty with sparse inverse co variances to couple multiple tasks. Learning multiple (parametric) models can be

viewed as estimating a matrix of parameters, where rows and columns of the matr
ix correspond to tasks and features, respectively. Following the matrix-variate
normal density, we design a penalty that decomposes the full covariance of matri
x elements into the Kronecker product of row covariance and column covariance, w
hich characterizes both task relatedness and feature representation. Several rec
ently proposed methods are variants of the special cases of this formulation. To
address the overfitting issue and select meaningful task and feature structures
, we include sparse covariance selection into our matrix-normal regularization v
ia L-1 penalties on task and feature inverse covariances. We empirically study t
he proposed method and compare with related models in two real-world problems: d
etecting landmines in multiple fields and recognizing faces between different su
bjects. Experimental results show that the proposed framework provides an effect
ive and flexible way to model various different structures of multiple tasks.
****

Network Flow Algorithms for Structured Sparsity
Julien Mairal, Rodolphe Jenatton, Francis Bach, Guillaume R. Obozinski
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Active Learning by Querying Informative and Representative Examples
Sheng-jun Huang, Rong Jin, Zhi-Hua Zhou
Most active learning approaches select either informative or representative unla
beled instances to query their labels. Although several active learning algorith
ms have been proposed to combine the two criterions for query selection, they ar
e usually ad hoc in finding unlabeled instances that are both informative and re
presentative. We address this challenge by a principled approach, termed QUIRE,
based on the min-max view of active learning. The proposed approach provides a s
ystematic way for measuring and combining the informativeness and representative
ness of an instance. Extensive experimental results show that the proposed QUIRE
approach outperforms several state-of -the-art active learning approaches.
****

Optimal Bayesian Recommendation Sets and Myopically Optimal Choice Query Sets
Paolo Viappiani, Craig Boutilier
Bayesian approaches to utility elicitation typically adopt (myopic) expected val
ue of information (EVOI) as a natural criterion for selecting queries. However,
EVOI-optimization is usually computationally prohibitive. In this paper, we e
xamine EVOI optimization using \emph{choice queries}, queries in which a user i
s ask to select her most preferred product from a set. We show that, under very
general assumptions, the optimal choice query w.r.t.\ EVOI coincides with \emph
{optimal recommendation set}, that is, a set maximizing expected utility of the
user selection. Since recommendation set optimization is a simpler, submodular p
roblem, this can greatly reduce the complexity of both exact and approximate (gr
eedy) computation of optimal choice queries. We also examine the case where us
er responses to choice queries are error-prone (using both constant and follow m
ixed multinomial logit noise models) and provide worst-case guarantees. Final
ly we present a local search technique that works well with large outcome spaces
.
****

Semi-Supervised Learning with Adversarially Missing Label Information
Umar Syed, Ben Taskar
We address the problem of semi-supervised learning in an adversarial setting. In
stead of assuming that labels are missing at random, we analyze a less favorable
scenario where the label information can be missing partially and arbitrarily,
which is motivated by several practical examples. We present nearly matching upp
er and lower generalization bounds for learning in this setting under reasonable
assumptions about available label information. Motivated by the analysis, we fo
rmulate a convex optimization problem for parameter estimation, derive an effici
ent algorithm, and analyze its convergence. We provide experimental results on s

everal standard data sets showing the robustness of our algorithm to the pattern of missing label information, outperforming several strong baselines.
****

Gated Softmax Classification
Roland Memisevic, Christopher Zach, Marc Pollefeys, Geoffrey E. Hinton
We describe a log-bilinear" model that computes class probabilities by combining an input vector multiplicatively with a vector of binary latent variables. Even though the latent variables can take on exponentially many possible combinations of values, we can efficiently compute the exact probability of each class by marginalizing over the latent variables. This makes it possible to get the exact gradient of the log likelihood. The bilinear score-functions are defined using a three-dimensional weight tensor, and we show that factorizing this tensor allows the model to encode invariances inherent in a task by learning a dictionary of invariant basis functions. Experiments on a set of benchmark problems show that this fully probabilistic model can achieve classification performance that is competitive with (kernel) SVMs, backpropagation, and deep belief nets."
****

Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs
Dávid Pál, Barnabás Póczos, Csaba Szepesvári
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Causal discovery in multiple models from different experiments
Tom Claassen, Tom Heskes
A long-standing open research problem is how to use information from different experiments, including background knowledge, to infer causal relations. Recent developments have shown ways to use multiple data sets, provided they originate from identical experiments. We present the MCI-algorithm as the first method that can infer provably valid causal relations in the large sample limit from different experiments. It is fast, reliable and produces very clear and easily interpretable output. It is based on a result that shows that constraint-based causal discovery is decomposable into a candidate pair identification and subsequent elimination step that can be applied separately from different models. We test the algorithm on a variety of synthetic input model sets to assess its behavior and the quality of the output. The method shows promising signs that it can be adapted to suit causal discovery in real-world application areas as well, including large databases.
****

Learning Bounds for Importance Weighting
Corinna Cortes, Yishay Mansour, Mehryar Mohri
This paper presents an analysis of importance weighting for learning from finite samples and gives a series of theoretical and algorithmic results. We point out simple cases where importance weighting can fail, which suggests the need for an analysis of the properties of this technique. We then give both upper and lower bounds for generalization with bounded importance weights and, more significantly, give learning guarantees for the more common case of unbounded importance weights under the weak assumption that the second moment is bounded, a condition related to the Renyi divergence of the training and test distributions. These results are based on a series of novel and general bounds we derive for unbounded loss functions, which are of independent interest. We use these bounds to guide the definition of an alternative reweighting algorithm and report the results of experiments demonstrating its benefits. Finally, we analyze the properties of normalized importance weights which are also commonly used.
****

A Reduction from Apprenticeship Learning to Classification
Umar Syed, Robert E. Schapire
Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Extensions of Generalized Binary Search to Group Identification and Exponential Costs
Gowtham Bellala, Suresh Bhavnani, Clayton Scott
Generalized Binary Search (GBS) is a well known greedy algorithm for identifying an unknown object while minimizing the number of yes" or "no" questions posed about that object, and arises in problems such as active learning and active diagnosis. Here, we provide a coding-theoretic interpretation for GBS and show that GBS can be viewed as a top-down algorithm that greedily minimizes the expected number of queries required to identify an object. This interpretation is then used to extend GBS in two ways. First, we consider the case where the objects are partitioned into groups, and the objective is to identify only the group to which the object belongs. Then, we consider the case where the cost of identifying an object grows exponentially in the number of queries. In each case, we present an exact formula for the objective function involving Shannon or Renyi entropy, and develop a greedy algorithm for minimizing it."
****

Feature Set Embedding for Incomplete Data
David Grangier, Iain Melvin
We present a new learning strategy for classification problems in which train and/or test data suffer from missing features. In previous work, instances are represented as vectors from some feature space and one is forced to impute missing values or to consider an instance-specific subspace. In contrast, our method considers instances as sets of (feature,value) pairs which naturally handle the missing value case. Building onto this framework, we propose a classification strategy for sets. Our proposal maps (feature,value) pairs into an embedding space and then non-linearly combines the set of embedded vectors. The embedding and the combination parameters are learned jointly on the final classification objective. This simple strategy allows great flexibility in encoding prior knowledge about the features in the embedding step and yields advantageous results compared to alternative solutions over several datasets.
****

Improving Human Judgments by Decontaminating Sequential Dependencies
Michael C. Mozer, Harold Pashler, Matthew Wilder, Robert V. Lindsey, Matt Jones, Michael N. Jones
For over half a century, psychologists have been struck by how poor people are at expressing their internal sensations, impressions, and evaluations via rating scales. When individuals make judgments, they are incapable of using an absolute rating scale, and instead rely on reference points from recent experience. This relativity of judgment limits the usefulness of responses provided by individuals to surveys, questionnaires, and evaluation forms. Fortunately, the cognitive processes that transform internal states to responses are not simply noisy, but rather are influenced by recent experience in a lawful manner. We explore techniques to remove sequential dependencies, and thereby decontaminate a series of ratings to obtain more meaningful human judgments. In our formulation, decontamination is fundamentally a problem of inferring latent states (internal sensations) which, because of the relativity of judgment, have temporal dependencies. We propose a decontamination solution using a conditional random field with constraints motivated by psychological theories of relative judgment. Our exploration of decontamination models is supported by two experiments we conducted to obtain ground-truth rating data on a simple length estimation task. Our decontamination techniques yield an over 20% reduction in the error of human judgments.
****

Scrambled Objects for Least-Squares Regression
Odalric Maillard, Rémi Munos
We consider least-squares regression using a randomly generated subspace GP\subset F of finite dimension P, where F is a function space of infinite dimension, e

.g.~L2([0,1]^d). GP is defined as the span of P random features that are linea r combinations of the basis functions of F weighted by random Gaussian i.i.d.~co efficients. In particular, we consider multi-resolution random combinations at a ll scales of a given mother function, such as a hat function or a wavelet. In t his latter case, the resulting Gaussian objects are called {\em scrambled wavele ts} and we show that they enable to approximate functions in Sobolev spaces H^s([0,1]^d). As a result, given N data, the least-squares estimate \hat g built fro m P scrambled wavelets has excess risk ||f^* - \hat g||\P^2 = O(||f^||^2_{H^s([0 ,1]^d)}(\log N)/P + P(\log N)/N) for target functions f^\in H^s([0,1]^d) of smo othness order s>d/2. An interesting aspect of the resulting bounds is that they do not depend on the distribution \P from which the data are generated, which is important in a statistical regression setting considered here. Randomization en ables to adapt to any possible distribution. We conclude by describing an effi cient numerical implementation using lazy expansions with numerical complexity \ tilde O(2^d N^{3/2}\log N + N^2), where d is the dimension of the input space.
****

Subgraph Detection Using Eigenvector L1 Norms
Benjamin Miller, Nadya Bliss, Patrick Wolfe
When working with network datasets, the theoretical framework of detection theor y for Euclidean vector spaces no longer applies. Nevertheless, it is desirable t o determine the detectability of small, anomalous graphs embedded into backgroun d networks with known statistical properties. Casting the problem of subgraph de tection in a signal processing context, this article provides a framework and em pirical results that elucidate a detection theory" for graph-valued data. Its fo cus is the detection of anomalies in unweighted, undirected graphs through L1 pr operties of the eigenvectors of the graph’s so-called modularity matrix. This me tric is observed to have relatively low variance for certain categories of rando mly-generated graphs, and to reveal the presence of an anomalous subgraph with r easonable reliability when the anomaly is not well-correlated with stronger port ions of the background graph. An analysis of subgraphs in real network datasets confirms the efficacy of this approach."
****

Error Propagation for Approximate Policy and Value Iteration
Amir-massoud Farahmand, Csaba Szepesvári, Rémi Munos
We address the question of how the approximation error/Bellman residual at each iteration of the Approximate Policy/Value Iteration algorithms influences the qu ality of the resulted policy. We quantify the performance loss as the Lp norm of the approximation error/Bellman residual at each iteration. Moreover, we show t hat the performance loss depends on the expectation of the squared Radon-Nikodym derivative of a certain distribution rather than its supremum -- as opposed to what has been suggested by the previous results. Also our results indicate that the contribution of the approximation/Bellman error to the performance loss is more prominent in the later iterations of API/AVI, and the effect of an error te rm in the earlier iterations decays exponentially fast.
****

PAC-Bayesian Model Selection for Reinforcement Learning
M. Fard, Joelle Pineau
This paper introduces the first set of PAC-Bayesian bounds for the batch reinfor cement learning problem in finite state spaces. These bounds hold regardless of the correctness of the prior distribution. We demonstrate how such bounds can be used for model-selection in control problems where prior information is availab le either on the dynamics of the environment, or on the value of actions. Our em pirical results confirm that PAC-Bayesian model-selection is able to leverage pr ior distributions when they are informative and, unlike standard Bayesian RL app roaches, ignores them when they are misleading.
****

Learning to combine foveal glimpses with a third-order Boltzmann machine
Hugo Larochelle, Geoffrey E. Hinton
We describe a model based on a Boltzmann machine with third-order connections th at can learn how to accumulate information about a shape over several fixations.

The model uses a retina that only has enough high resolution pixels to cover a small area of the image, so it must decide on a sequence of fixations and it must combine the glimpse" at each fixation with the location of the fixation before integrating the information with information from other glimpses of the same object. We evaluate this model on a synthetic dataset and two image classification datasets, showing that it can perform at least as well as a model trained on whole images."

****

Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm

Nathan Srebro, Russ R. Salakhutdinov

We show that matrix completion with trace-norm regularization can be significantly hurt when entries of the matrix are sampled non-uniformly, but that a properly weighted version of the trace-norm regularizer works well with non-uniform sampling. We show that the weighted trace-norm regularization indeed yields significant gains on the highly non-uniformly sampled Netflix dataset.

****

Worst-case bounds on the quality of max-product fixed-points

Meritxell Vinyals, Jes\'us Cerquides, Alessandro Farinelli, Juan Rodríguez-aguilar

We study worst-case bounds on the quality of any fixed point assignment of the max-product algorithm for Markov Random Fields (MRF). We start proving a bound independent of the MRF structure and parameters. Afterwards, we show how this bound can be improved for MRFs with particular structures such as bipartite graphs or grids. Our results provide interesting insight into the behavior of max-product. For example, we prove that max-product provides very good results (at least 90% of the optimal) on MRFs with large variable-disjoint cycles (MRFs in which all cycles are variable-disjoint, namely that they do not share any edge and in which each cycle contains at least 20 variables).

****

A POMDP Extension with Belief-dependent Rewards

Mauricio Araya, Olivier Buffet, Vincent Thomas, Françcois Charpillet

Partially Observable Markov Decision Processes (POMDPs) model sequential decision-making problems under uncertainty and partial observability. Unfortunately, some problems cannot be modeled with state-dependent reward functions, e.g., problems whose objective explicitly implies reducing the uncertainty on the state. To that end, we introduce rho-POMDPs, an extension of POMDPs where the reward function rho depends on the belief state. We show that, under the common assumption that rho is convex, the value function is also convex, what makes it possible to (1) approximate rho arbitrarily well with a piecewise linear and convex (PWLC) function, and (2) use state-of-the-art exact or approximate solving algorithms with limited changes.

****

Infinite Relational Modeling of Functional Connectivity in Resting State fMRI

Morten Mørup, Kristoffer Madsen, Anne-marie Dogonowski, Hartwig Siebner, Lars K. Hansen

Functional magnetic resonance imaging (fMRI) can be applied to study the functional connectivity of the neural elements which form complex network at a whole brain level. Most analyses of functional resting state networks (RSN) have been based on the analysis of correlation between the temporal dynamics of various regions of the brain. While these models can identify coherently behaving groups in terms of correlation they give little insight into how these groups interact. In this paper we take a different view on the analysis of functional resting state networks. Starting from the definition of resting state as functional coherent groups we search for functional units of the brain that communicate with other parts of the brain in a coherent manner as measured by mutual information. We use the infinite relational model (IRM) to quantify functional coherent groups of resting state networks and demonstrate how the extracted component interactions can be used to discriminate between functional resting state activity in multiple sclerosis and normal subjects.

****
An Alternative to Low-level-Sychrony-Based Methods for Speech Detection
Javier Movellan, Paul Ruvolo
Determining whether someone is talking has applications in many areas such as speech recognition, speaker diarization, social robotics, facial expression recognition, and human computer interaction. One popular approach to this problem is audio-visual synchrony detection. A candidate speaker is deemed to be talking if the visual signal around that speaker correlates with the auditory signal. Here we show that with the proper visual features (in this case movements of various facial muscle groups), a very accurate detector of speech can be created that does not use the audio signal at all. Further we show that this person independent visual-only detector can be used to train very accurate audio-based person dependent voice models. The voice model has the advantage of being able to identify when a particular person is speaking even when they are not visible to the camera (e.g. in the case of a mobile robot). Moreover, we show that a simple sensory fusion scheme between the auditory and visual models improves performance on the task of talking detection. The work here provides dramatic evidence about the efficacy of two very different approaches to multimodal speech detection on a challenging database.
****
A rational decision making framework for inhibitory control
Pradeep Shenoy, Angela J. Yu, Rajesh PN Rao
Intelligent agents are often faced with the need to choose actions with uncertain consequences, and to modify those actions according to ongoing sensory processing and changing task demands. The requisite ability to dynamically modify or cancel planned actions is known as inhibitory control in psychology. We formalize inhibitory control as a rational decision-making problem, and apply to it to the classical stop-signal task. Using Bayesian inference and stochastic control tools, we show that the optimal policy systematically depends on various parameters of the problem, such as the relative costs of different action choices, the noise level of sensory inputs, and the dynamics of changing environmental demands. Our normative model accounts for a range of behavioral data in humans and animals in the stop-signal task, suggesting that the brain implements statistically optimal, dynamically adaptive, and reward-sensitive decision-making in the context of inhibitory control problems.
****
Policy gradients in linearly-solvable MDPs
Emanuel Todorov
We present policy gradient results within the framework of linearly-solvable MDPs. For the first time, compatible function approximators and natural policy gradients are obtained by estimating the cost-to-go function, rather than the (much larger) state-action advantage function as is necessary in traditional MDPs. We also develop the first compatible function approximators and natural policy gradients for continuous-time stochastic systems.
****
Sufficient Conditions for Generating Group Level Sparsity in a Robust Minimax Framework
Hongbo Zhou, Qiang Cheng
Regularization technique has become a principle tool for statistics and machine learning research and practice. However, in most situations, these regularization terms are not well interpreted, especially on how they are related to the loss function and data. In this paper, we propose a robust minimax framework to interpret the relationship between data and regularization terms for a large class of loss functions. We show that various regularization terms are essentially corresponding to different distortions to the original data matrix. This minimax framework includes ridge regression, lasso, elastic net, fused lasso, group lasso, local coordinate coding, multiple kernel learning, etc., as special cases. Within this minimax framework, we further gave mathematically exact definition for a novel representation called sparse grouping representation (SGR), and proved sufficient conditions for generating such group level sparsity. Under these suffici

ent conditions, a large set of consistent regularization terms can be designed. This SGR is essentially different from group lasso in the way of using class or group information, and it outperforms group lasso when there appears group label noise. We also gave out some generalization bounds in a classification setting.
****

Learning concept graphs from text with stick-breaking priors
America Chambers, Padhraic Smyth, Mark Steyvers
We present a generative probabilistic model for learning general graph structures, which we term concept graphs, from text. Concept graphs provide a visual summary of the thematic content of a collection of documents-a task that is difficult to accomplish using only keyword search. The proposed model can learn different types of concept graph structures and is capable of utilizing partial prior knowledge about graph structure as well as labeled documents. We describe a generative model that is based on a stick-breaking process for graphs, and a Markov Chain Monte Carlo inference procedure. Experiments on simulated data show that the model can recover known graph structure when learning in both unsupervised and semi-supervised modes. We also show that the proposed model is competitive in terms of empirical log likelihood with existing structure-based topic models (such as hPAM and hLDA) on real-world text data sets. Finally, we illustrate the application of the model to the problem of updating Wikipedia category graphs.
****

Fast Large-scale Mixture Modeling with Component-specific Data Partitions
Bo Thiesson, Chong Wang
Remarkably easy implementation and guaranteed convergence has made the EM algorithm one of the most used algorithms for mixture modeling. On the downside, the E-step is linear in both the sample size and the number of mixture components, making it impractical for large-scale data. Based on the variational EM framework, we propose a fast alternative that uses component-specific data partitions to obtain a sub-linear E-step in sample size, while the algorithm still maintains provable convergence. Our approach builds on previous work, but is significantly faster and scales much better in the number of mixture components. We demonstrate this speedup by experiments on large-scale synthetic and real data.
****

Improvements to the Sequence Memoizer
Jan Gasthaus, Yee Teh
The sequence memoizer is a model for sequence data with state-of-the-art performance on language modeling and compression. We propose a number of improvements to the model and inference algorithm, including an enlarged range of hyperparameters, a memory-efficient representation, and inference algorithms operating on the new representation. Our derivations are based on precise definitions of the various processes that will also allow us to provide an elementary proof of the mysterious" coagulation and fragmentation properties used in the original paper on the sequence memoizer by Wood et al. (2009). We present some experimental results supporting our improvements."
****

Simultaneous Object Detection and Ranking with Weak Supervision
Matthew Blaschko, Andrea Vedaldi, Andrew Zisserman
A standard approach to learning object category detectors is to provide strong supervision in the form of a region of interest (ROI) specifying each instance of the object in the training images. In this work are goal is to learn from heterogeneous labels, in which some images are only weakly supervised, specifying only the presence or absence of the object or a weak indication of object location, whilst others are fully annotated. To this end we develop a discriminative learning approach and make two contributions: (i) we propose a structured output formulation for weakly annotated images where full annotations are treated as latent variables; and (ii) we propose to optimize a ranking objective function, allowing our method to more effectively use negatively labeled images to improve detection average precision performance. The method is demonstrated on the benchmark INRIA pedestrian detection dataset of Dalal and Triggs and the PASCAL VOC dataset, and it is shown that for a significant proportion of weakly supervised image

s the performance achieved is very similar to the fully supervised (state of the art) results.
****

Generating more realistic images using gated MRF's
Marc'aurelio Ranzato, Volodymyr Mnih, Geoffrey E. Hinton
Probabilistic models of natural images are usually evaluated by measuring performance on rather indirect tasks, such as denoising and inpainting. A more direct way to evaluate a generative model is to draw samples from it and to check whether statistical properties of the samples match the statistics of natural images. This method is seldom used with high-resolution images, because current models produce samples that are very different from natural images, as assessed by even simple visual inspection. We investigate the reasons for this failure and we show that by augmenting existing models so that there are two sets of latent variables, one set modelling pixel intensities and the other set modelling image-specific pixel covariances, we are able to generate high-resolution images that look much more realistic than before. The overall model can be interpreted as a gated MRF where both pair-wise dependencies and mean intensities of pixels are modulated by the states of latent variables. Finally, we confirm that if we disallow weight-sharing between receptive fields that overlap each other, the gated MRF learns more efficient internal representations, as demonstrated in several recognition tasks.
****

Efficient Minimization of Decomposable Submodular Functions
Peter Stobbe, Andreas Krause
Many combinatorial problems arising in machine learning can be reduced to the problem of minimizing a submodular function. Submodular functions are a natural discrete analog of convex functions, and can be minimized in strongly polynomial time. Unfortunately, state-of-the-art algorithms for general submodular minimization are intractable for practical problems. In this paper, we introduce a novel subclass of submodular minimization problems that we call decomposable. Decomposable submodular functions are those that can be represented as sums of concave functions applied to linear functions. We develop an algorithm, SLG, that can efficiently minimize decomposable submodular functions with tens of thousands of variables. Our algorithm exploits recent results in smoothed convex minimization. We apply SLG to synthetic benchmarks and a joint classification-and-segmentation task, and show that it outperforms the state-of-the-art general purpose submodular minimization algorithms by several orders of magnitude.
****

Implicit Differentiation by Perturbation
Justin Domke
This paper proposes a simple and efficient finite difference method for implicit differentiation of marginal inference results in discrete graphical models. Given an arbitrary loss function, defined on marginals, we show that the derivatives of this loss with respect to model parameters can be obtained by running the inference procedure twice, on slightly perturbed model parameters. This method can be used with approximate inference, with a loss function over approximate marginals. Convenient choices of loss functions make it practical to fit graphical models with hidden variables, high treewidth and/or model misspecification.
****

The Maximal Causes of Natural Scenes are Edge Filters
Jose Puertas, Joerg Bornschein, Jörg Lücke
We study the application of a strongly non-linear generative model to image patches. As in standard approaches such as Sparse Coding or Independent Component Analysis, the model assumes a sparse prior with independent hidden variables. However, in the place where standard approaches use the sum to combine basis functions we use the maximum. To derive tractable approximations for parameter estimation we apply a novel approach based on variational Expectation Maximization. The derived learning algorithm can be applied to large-scale problems with hundreds of observed and hidden variables. Furthermore, we can infer all model parameters including observation noise and the degree of sparseness. In applications to im

age patches we find that Gabor-like basis functions are obtained. Gabor-like fun ctions are thus not a feature exclusive to approaches assuming linear superposit ion. Quantitatively, the inferred basis functions show a large diversity of shap es with many strongly elongated and many circular symmetric functions. The distr ibution of basis function shapes reflects properties of simple cell receptive fi elds that are not reproduced by standard linear approaches. In the study of natu ral image statistics, the implications of using different superposition assumpti ons have so far not been investigated systematically because models with strong non-linearities have been found analytically and computationally challenging. Th e presented algorithm represents the first large-scale application of such an ap proach.

****

Regularized estimation of image statistics by Score Matching

Durk P. Kingma, Yann Cun

Score Matching is a recently-proposed criterion for training high-dimensional de nsity models for which maximum likelihood training is intractable. It has been a pplied to learning natural image statistics but has so-far been limited to simpl e models due to the difficulty of differentiating the loss with respect to the m odel parameters. We show how this differentiation can be automated with an exten ded version of the double-backpropagation algorithm. In addition, we introduce a regularization term for the Score Matching loss that enables its use for a broa der range of problem by suppressing instabilities that occur with finite trainin g sample sizes and quantized input values. Results are reported for image denois ing and super-resolution.

****

Identifying Patients at Risk of Major Adverse Cardiovascular Events Using Symbol ic Mismatch

Zeeshan Syed, John Guttag

Cardiovascular disease is the leading cause of death globally, resulting in 17 m illion deaths each year. Despite the availability of various treatment options, existing techniques based upon conventional medical knowledge often fail to iden tify patients who might have benefited from more aggressive therapy. In this pap er, we describe and evaluate a novel unsupervised machine learning approach for cardiac risk stratification. The key idea of our approach is to avoid specialize d medical knowledge, and assess patient risk using symbolic mismatch, a new metr ic to assess similarity in long-term time-series activity. We hypothesize that h igh risk patients can be identified using symbolic mismatch, as individuals in a population with unusual long-term physiological activity. We describe related a pproaches that build on these ideas to provide improved medical decision making for patients who have recently suffered coronary attacks. We first describe how to compute the symbolic mismatch between pairs of long term electrocardiographic (ECG) signals. This algorithm maps the original signals into a symbolic domain, and provides a quantitative assessment of the difference between these symbolic representations of the original signals. We then show how this measure can be u sed with each of a one-class SVM, a nearest neighbor classifier, and hierarchica l clustering to improve risk stratification. We evaluated our methods on a popul ation of 686 cardiac patients with available long-term electrocardiographic data . In a univariate analysis, all of the methods provided a statistically signific ant association with the occurrence of a major adverse cardiac event in the next 90 days. In a multivariate analysis that incorporated the most widely used clin ical risk variables, the nearest neighbor and hierarchical clustering approaches were able to statistically significantly distinguish patients with a roughly tw o-fold risk of suffering a major adverse cardiac event in the next 90 days.

****

Movement extraction by detecting dynamics switches and repetitions

Silvia Chiappa, Jan Peters

Many time-series such as human movement data consist of a sequence of basic acti ons, e.g., forehands and backhands in tennis. Automatically extracting and chara cterizing such actions is an important problem for a variety of different applic ations. In this paper, we present a probabilistic segmentation approach in which

an observed time-series is modeled as a concatenation of segments corresponding to different basic actions. Each segment is generated through a noisy transformation of one of a few hidden trajectories representing different types of movement, with possible time re-scaling. We analyze three different approximation methods for dealing with model intractability, and demonstrate how the proposed approach can successfully segment table tennis movements recorded using a robot arm as haptic input device.
****
Exact inference and learning for cumulative distribution functions on loopy graphs
Nebojsa Jojic, Chris Meek, Jim Huang
Probabilistic graphical models use local factors to represent dependence among sets of variables. For many problem domains, for instance climatology and epidemiology, in addition to local dependencies, we may also wish to model heavy-tailed statistics, where extreme deviations should not be treated as outliers. Specifying such distributions using graphical models for probability density functions (PDFs) generally lead to intractable inference and learning. Cumulative distribution networks (CDNs) provide a means to tractably specify multivariate heavy-tailed models as a product of cumulative distribution functions (CDFs). Currently, algorithms for inference and learning, which correspond to computing mixed derivatives, are exact only for tree-structured graphs. For graphs of arbitrary topology, an efficient algorithm is needed that takes advantage of the sparse structure of the model, unlike symbolic differentiation programs such as Mathematica and D* that do not. We present an algorithm for recursively decomposing the computation of derivatives for CDNs of arbitrary topology, where the decomposition is naturally described using junction trees. We compare the performance of the resulting algorithm to Mathematica and D*, and we apply our method to learning models for rainfall and H1N1 data, where we show that CDNs with cycles are able to provide a significantly better fits to the data as compared to tree-structured and unstructured CDNs and other heavy-tailed multivariate distributions such as the multivariate copula and logistic models.
****
Spectral Regularization for Support Estimation
Ernesto Vito, Lorenzo Rosasco, Alessandro Toigo
In this paper we consider the problem of learning from data the support of a probability distribution when the distribution {\em does not} have a density (with respect to some reference measure). We propose a new class of regularized spectral estimators based on a new notion of reproducing kernel Hilbert space, which we call {\em ``completely regular''}. Completely regular kernels allow to capture the relevant geometric and topological properties of an arbitrary probability space. In particular, they are the key ingredient to prove the universal consistency of the spectral estimators and in this respect they are the analogue of universal kernels for supervised problems. Numerical experiments show that spectral estimators compare favorably to state of the art machine learning algorithms for density support estimation.
****
Online Learning for Latent Dirichlet Allocation
Matthew Hoffman, Francis Bach, David Blei
We develop an online variational Bayes (VB) algorithm for Latent Dirichlet Allocation (LDA). Online LDA is based on online stochastic optimization with a natural gradient step, which we show converges to a local optimum of the VB objective function. It can handily analyze massive document collections, including those arriving in a stream. We study the performance of online LDA in several ways, including by fitting a 100-topic topic model to 3.3M articles from Wikipedia in a single pass. We demonstrate that online LDA finds topic models as good or better than those found with batch VB, and in a fraction of the time.
****
Random Projections for k-means Clustering
Christos Boutsidis, Anastasios Zouzias, Petros Drineas
Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Inference and communication in the game of Password
Yang Xu, Charles Kemp
Communication between a speaker and hearer will be most efficient when both parties make accurate inferences about the other. We study inference and communication in a television game called Password, where speakers must convey secret words to hearers by providing one-word clues. Our working hypothesis is that human communication is relatively efficient, and we use game show data to examine three predictions. First, we predict that speakers and hearers are both considerate, and that both take the other's perspective into account. Second, we predict that speakers and hearers are calibrated, and that both make accurate assumptions about the strategy used by the other. Finally, we predict that speakers and hearers are collaborative, and that they tend to share the cognitive burden of communication equally. We find evidence in support of all three predictions, and demonstrate in addition that efficient communication tends to break down when speakers and hearers are placed under time pressure.
****

Smoothness, Low Noise and Fast Rates
Nathan Srebro, Karthik Sridharan, Ambuj Tewari
We establish an excess risk bound of O(H Rn^2 + sqrt{H L*} Rn) for ERM with an H-smooth loss function and a hypothesis class with Rademacher complexity Rn, where L* is the best risk achievable by the hypothesis class. For typical hypothesis classes where Rn = sqrt{R/n}, this translates to a learning rate of ■ O(RH/n) in the separable (L* = 0) case and O(RH/n + sqrt{L* RH/n}) more generally. We also provide similar guarantees for online and stochastic convex optimization of a smooth non-negative objective.
****

Energy Disaggregation via Discriminative Sparse Coding
J. Kolter, Siddharth Batra, Andrew Ng
Energy disaggregation is the task of taking a whole-home energy signal and separating it into its component appliances. Studies have shown that having device-level energy information can cause users to conserve significant amounts of energy, but current electricity meters only report whole-home data. Thus, developing algorithmic methods for disaggregation presents a key technical challenge in the effort to maximize energy conservation. In this paper, we examine a large scale energy disaggregation task, and apply a novel extension of sparse coding to this problem. In particular, we develop a method, based upon structured prediction, for discriminatively training sparse coding algorithms specifically to maximize disaggregation performance. We show that this significantly improves the performance of sparse coding algorithms on the energy task and illustrate how these disaggregation results can provide useful information about energy usage.
****

Random Conic Pursuit for Semidefinite Programming
Ariel Kleiner, Ali Rahimi, Michael Jordan
We present a novel algorithm, Random Conic Pursuit, that solves semidefinite programs (SDPs) via repeated optimization over randomly selected two-dimensional subcones of the PSD cone. This scheme is simple, easily implemented, applicable to very general SDPs, scalable, and theoretically interesting. Its advantages are realized at the expense of an ability to readily compute highly exact solutions, though useful approximate solutions are easily obtained. This property renders Random Conic Pursuit of particular interest for machine learning applications, in which the relevant SDPs are generally based upon random data and so exact minima are often not a priority. Indeed, we present empirical results to this effect for various SDPs encountered in machine learning; these experiments demonstrate the potential practical usefulness of Random Conic Pursuit. We also provide a preliminary analysis that yields insight into the theoretical properties and convergence of the algorithm.

****
Deterministic Single-Pass Algorithm for LDA
Issei Sato, Kenichi Kurihara, Hiroshi Nakagawa
We develop a deterministic single-pass algorithm for latent Dirichlet allocation (LDA) in order to process received documents one at a time and then discard them in an excess text stream. Our algorithm does not need to store old statistics for all data. The proposed algorithm is much faster than a batch algorithm and is comparable to the batch algorithm in terms of perplexity in experiments.
****
A Bayesian Framework for Figure-Ground Interpretation
Vicky Froyen, Jacob Feldman, Manish Singh
Figure/ground assignment, in which the visual image is divided into nearer (figural) and farther (ground) surfaces, is an essential step in visual processing, but its underlying computational mechanisms are poorly understood. Figural assignment (often referred to as border ownership) can vary along a contour, suggesting a spatially distributed process whereby local and global cues are combined to yield local estimates of border ownership. In this paper we model figure/ground estimation in a Bayesian belief network, attempting to capture the propagation of border ownership across the image as local cues (contour curvature and T-junctions) interact with more global cues to yield a figure/ground assignment. Our network includes as a nonlocal factor skeletal (medial axis) structure, under the hypothesis that medial structure ``draws'' border ownership so that borders are owned by their interiors. We also briefly present a psychophysical experiment in which we measured local border ownership along a contour at various distances from an inducing cue (a T-junction). Both the human subjects and the network show similar patterns of performance, converging rapidly to a similar pattern of spatial variation in border ownership along contours.
****
Online Markov Decision Processes under Bandit Feedback
Gergely Neu, Andras Antos, András György, Csaba Szepesvári
We consider online learning in finite stochastic Markovian environments where in each time step a new reward function is chosen by an oblivious adversary. The goal of the learning agent is to compete with the best stationary policy in terms of the total reward received. In each time step the agent observes the current state and the reward associated with the last transition, however, the agent does not observe the rewards associated with other state-action pairs. The agent is assumed to know the transition probabilities. The state of the art result for this setting is a no-regret algorithm. In this paper we propose a new learning algorithm and assuming that stationary policies mix uniformly fast, we show that after T time steps, the expected regret of the new algorithm is O(T^{2/3} (ln T)^{1/3}), giving the first rigorously proved convergence rate result for the problem.
****
A Log-Domain Implementation of the Diffusion Network in Very Large Scale Integration
Yi-da Wu, Shi-jie Lin, Hsin Chen
The Diffusion Network(DN) is a stochastic recurrent network which has been shown capable of modeling the distributions of continuous-valued, continuous-time paths. However, the dynamics of the DN are governed by stochastic differential equations, making the DN unfavourable for simulation in a digital computer. This paper presents the implementation of the DN in analogue Very Large Scale Integration, enabling the DN to be simulated in real time. Moreover, the log-domain representation is applied to the DN, allowing the supply voltage and thus the power consumption to be reduced without limiting the dynamic ranges for diffusion processes. A VLSI chip containing a DN with two stochastic units has been designed and fabricated. The design of component circuits will be described, so will the simulation of the full system be presented. The simulation results demonstrate that the DN in VLSI is able to regenerate various types of continuous paths in real-time.
****

SpikeAnts, a spiking neuron network modelling the emergence of organization in a complex system
Sylvain Chevallier, Hél\`ene Paugam-moisy, Michele Sebag
Many complex systems, ranging from neural cell assemblies to insect societies, involve and rely on some division of labor. How to enforce such a division in a decentralized and distributed way, is tackled in this paper, using a spiking neuron network architecture. Specifically, a spatio-temporal model called SpikeAnts is shown to enforce the emergence of synchronized activities in an ant colony. Each ant is modelled from two spiking neurons; the ant colony is a sparsely connected spiking neuron network. Each ant makes its decision (among foraging, sleeping and self-grooming) from the competition between its two neurons, after the signals received from its neighbor ants. Interestingly, three types of temporal patterns emerge in the ant colony: asynchronous, synchronous, and synchronous periodic foraging activities - similar to the actual behavior of some living ant colonies. A phase diagram of the emergent activity patterns with respect to two control parameters, respectively accounting for ant sociability and receptivity, is presented and discussed.
****
Permutation Complexity Bound on Out-Sample Error
Malik Magdon-Ismail
We define a data dependent permutation complexity for a hypothesis set \math{\hset}, which is similar to a Rademacher complexity or maximum discrepancy. The permutation complexity is based like the maximum discrepancy on (dependent) sampling. We prove a uniform bound on the generalization error, as well as a concentration result which means that the permutation estimate can be efficiently estimated.
****
Fast global convergence rates of gradient methods for high-dimensional statistical recovery
Alekh Agarwal, Sahand Negahban, Martin J. Wainwright
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Attractor Dynamics with Synaptic Depression
K. Wong, He Wang, Si Wu, Chi Fung
Neuronal connection weights exhibit short-term depression (STD). The present study investigates the impact of STD on the dynamics of a continuous attractor neural network (CANN) and its potential roles in neural information processing. We find that the network with STD can generate both static and traveling bumps, and STD enhances the performance of the network in tracking external inputs. In particular, we find that STD endows the network with slow-decaying plateau behaviors, namely, the network being initially stimulated to an active state will decay to silence very slowly in the time scale of STD rather than that of neural signaling. We argue that this provides a mechanism for neural systems to hold short-term memory easily and shut off persistent activities naturally.
****
Layer-wise analysis of deep networks with Gaussian kernels
Grégoire Montavon, Klaus-Robert Müller, Mikio Braun
Deep networks can potentially express a learning problem more efficiently than local learning machines. While deep networks outperform local learning machines on some problems, it is still unclear how their nice representation emerges from their complex structure. We present an analysis based on Gaussian kernels that measures how the representation of the learning problem evolves layer after layer as the deep network builds higher-level abstract representations of the input. We use this analysis to show empirically that deep networks build progressively better representations of the learning problem and that the best representations are obtained when the deep network discriminates only in the last layers.
****

Space-Variant Single-Image Blind Deconvolution for Removing Camera Shake
Stefan Harmeling, Hirsch Michael, Bernhard Schölkopf
Modelling camera shake as a space-invariant convolution simplifies the problem o f removing camera shake, but often insufficiently models actual motion blur such as those due to camera rotation and movements outside the sensor plane or when objects in the scene have different distances to the camera. In order to overcom e such limitations we contribute threefold: (i) we introduce a taxonomy of camer a shakes, (ii) we show how to combine a recently introduced framework for space-variant filtering based on overlap-add from Hirsch et al.~and a fast algorithm f or single image blind deconvolution for space-invariant filters from Cho and Lee to introduce a method for blind deconvolution for space-variant blur. And (iii) , we present an experimental setup for evaluation that allows us to take images with real camera shake while at the same time record the space-variant point spr ead function corresponding to that blur. Finally, we demonstrate that our method is able to deblur images degraded by spatially-varying blur originating from re al camera shake.
****
Sodium entry efficiency during action potentials: A novel single-parameter famil y of Hodgkin-Huxley models
Anand Singh, Renaud Jolivet, Pierre Magistretti, Bruno Weber
Sodium entry during an action potential determines the energy efficiency of a ne uron. The classic Hodgkin-Huxley model of action potential generation is notorio usly inefficient in that regard with about 4 times more charges flowing through the membrane than the theoretical minimum required to achieve the observed depol arization. Yet, recent experimental results show that mammalian neurons are clos e to the optimal metabolic efficiency and that the dynamics of their voltage-gat ed channels is significantly different than the one exhibited by the classic Hod gkin-Huxley model during the action potential. Nevertheless, the original Hodgki n-Huxley model is still widely used and rarely to model the squid giant axon fro m which it was extracted. Here, we introduce a novel family of Hodgkin-Huxley mo dels that correctly account for sodium entry, action potential width and whose v oltage-gated channels display a dynamics very similar to the most recent experim ental observations in mammalian neurons. We speak here about a family of models because the model is parameterized by a unique parameter the variations of which allow to reproduce the entire range of experimental observations from cortical pyramidal neurons to Purkinje cells, yielding a very economical framework to mod el a wide range of different central neurons. The present paper demonstrates the performances and discuss the properties of this new family of models.
****
Global Analytic Solution for Variational Bayesian Matrix Factorization
Shinichi Nakajima, Masashi Sugiyama, Ryota Tomioka
Bayesian methods of matrix factorization (MF) have been actively explored recent ly as promising alternatives to classical singular value decomposition. In this paper, we show that, despite the fact that the optimization problem is non-conve x, the global optimal solution of variational Bayesian (VB) MF can be computed a nalytically by solving a quartic equation. This is highly advantageous over a po pular VBMF algorithm based on iterated conditional modes since it can only find a local optimal solution after iterations. We further show that the global optim al solution of empirical VBMF (hyperparameters are also learned from data) can a lso be analytically computed. We illustrate the usefulness of our results throug h experiments.
****
Improving the Asymptotic Performance of Markov Chain Monte-Carlo by Inserting Vo rtices
Yi Sun, Jürgen Schmidhuber, Faustino Gomez
We present a new way of converting a reversible finite Markov chain into a nonre versible one, with a theoretical guarantee that the asymptotic variance of the M CMC estimator based on the non-reversible chain is reduced. The method is applic able to any reversible chain whose states are not connected through a tree, and can be interpreted graphically as inserting vortices into the state transition g

raph. Our result confirms that non-reversible chains are fundamentally better th
an reversible ones in terms of asymptotic performance, and suggests interesting
directions for further improving MCMC.
****

Linear Complementarity for Regularized Policy Evaluation and Improvement
Jeffrey Johns, Christopher Painter-wakefield, Ronald Parr
Recent work in reinforcement learning has emphasized the power of L1 regularizat
ion to perform feature selection and prevent overfitting. We propose formulating
 the L1 regularized linear fixed point problem as a linear complementarity probl
em (LCP). This formulation offers several advantages over the LARS-inspired form
ulation, LARS-TD. The LCP formulation allows the use of efficient off-the-shelf
solvers, leads to a new uniqueness result, and can be initialized with starting
points from similar problems (warm starts). We demonstrate that warm starts, as
well as the efficiency of LCP solvers, can speed up policy iteration. Moreover,
warm starts permit a form of modified policy iteration that can be used to appro
ximate a greedy" homotopy path, a generalization of the LARS-TD homotopy path th
at combines policy evaluation and optimization."
****

Inter-time segment information sharing for non-homogeneous dynamic Bayesian netw
orks
Dirk Husmeier, Frank Dondelinger, Sophie Lebre
Conventional dynamic Bayesian networks (DBNs) are based on the homogeneous Marko
v assumption, which is too restrictive in many practical applications. Various a
pproaches to relax the homogeneity assumption have therefore been proposed in th
e last few years. The present paper aims to improve the flexibility of two recen
t versions of non-homogeneous DBNs, which either (i) suffer from the need for da
ta discretization, or (ii) assume a time-invariant network structure. Allowing t
he network structure to be fully flexible leads to the risk of overfitting and i
nflated inference uncertainty though, especially in the highly topical field of
systems biology, where independent measurements tend to be sparse. In the presen
t paper we investigate three conceptually different regularization schemes based
 on inter-segment information sharing. We assess the performance in a comparativ
e evaluation study based on simulated data. We compare the predicted segmentatio
n of gene expression time series obtained during embryogenesis in Drosophila mel
anogaster with other state-of-the-art techniques. We conclude our evaluation wit
h an application to synthetic biology, where the objective is to predict a known
 regulatory network of five genes in Saccharomyces cerevisiae.
****

Graph-Valued Regression
Han Liu, Xi Chen, Larry Wasserman, John Lafferty
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Probabilistic Multi-Task Feature Selection
Yu Zhang, Dit-Yan Yeung, Qian Xu
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Penalized Principal Component Regression on Graphs for Analysis of Subnetworks
Ali Shojaie, George Michailidis
Network models are widely used to capture interactions among component of comple
x systems, such as social and biological. To understand their behavior, it is of
ten necessary to analyze functionally related components of the system, correspo
nding to subsystems. Therefore, the analysis of subnetworks may provide addition
al insight into the behavior of the system, not evident from individual componen
ts. We propose a novel approach for incorporating available network information

into the analysis of arbitrary subnetworks. The proposed method offers an effici ent dimension reduction strategy using Laplacian eigenmaps with Neumann boundary conditions, and provides a flexible inference framework for analysis of subnetw orks, based on a group-penalized principal component regression model on graphs. Asymptotic properties of the proposed inference method, as well as the choice o f the tuning parameter for control of the false positive rate are discussed in h igh dimensional settings. The performance of the proposed methodology is illustr ated using simulated and real data examples from biology.

****

Bayesian Action-Graph Games

Albert Jiang, Kevin Leyton-brown

Games of incomplete information, or Bayesian games, are an important game-theore tic model and have many applications in economics. We propose Bayesian action-gr aph games (BAGGs), a novel graphical representation for Bayesian games. BAGGs ca n represent arbitrary Bayesian games, and furthermore can compactly express Baye sian games exhibiting commonly encountered types of structure including symmetry , action- and type-specific utility independence, and probabilistic independence of type distributions. We provide an algorithm for computing expected utility i n BAGGs, and discuss conditions under which the algorithm runs in polynomial tim e. Bayes-Nash equilibria of BAGGs can be computed by adapting existing algorith ms for complete-information normal form games and leveraging our expected utilit y algorithm. We show both theoretically and empirically that our approaches impr ove significantly on the state of the art.

****

A Family of Penalty Functions for Structured Sparsity

Jean Morales, Charles Micchelli, Massimiliano Pontil

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

****

Spike timing-dependent plasticity as dynamic filter

Joscha Schmiedt, Christian Albers, Klaus Pawelzik

When stimulated with complex action potential sequences synapses exhibit spike t iming-dependent plasticity (STDP) with attenuated and enhanced pre- and postsyna ptic contributions to long-term synaptic modifications. In order to investigate the functional consequences of these contribution dynamics (CD) we propose a min imal model formulated in terms of differential equations. We find that our model reproduces a wide range of experimental results with a small number of biophysi cally interpretable parameters. The model allows to investigate the susceptibili ty of STDP to arbitrary time courses of pre- and postsynaptic activities, i.e. i ts nonlinear filter properties. We demonstrate this for the simple example of sm all periodic modulations of pre- and postsynaptic firing rates for which our mod el can be solved. It predicts synaptic strengthening for synchronous rate modula tions. For low baseline rates modifications are dominant in the theta frequency range, a result which underlines the well known relevance of theta activities in hippocampus and cortex for learning. We also find emphasis of low baseline spik e rates and suppression for high baseline rates. The latter suggests a mechanism of network activity regulation inherent in STDP. Furthermore, our novel formula tion provides a general framework for investigating the joint dynamics of neuron al activity and the CD of STDP in both spike-based as well as rate-based neurona l network models.

****

The Neural Costs of Optimal Control

Samuel Gershman, Robert Wilson

Optimal control entails combining probabilities and utilities. However, for most practical problems probability densities can be represented only approximately. Choosing an approximation requires balancing the benefits of an accurate approx imation against the costs of computing it. We propose a variational framework fo r achieving this balance and apply it to the problem of how a population code sh

ould optimally represent a distribution under resource constraints. The essence of our analysis is the conjecture that population codes are organized to maximize a lower bound on the log expected utility. This theory can account for a plethora of experimental data, including the reward-modulation of sensory receptive fields.
****

Sample Complexity of Testing the Manifold Hypothesis
Hariharan Narayanan, Sanjoy Mitter
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

A biologically plausible network for the computation of orientation dominance
Kritika Muralidharan, Nuno Vasconcelos
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

A Primal-Dual Algorithm for Group Sparse Regularization with Overlapping Groups
Sofia Mosci, Silvia Villa, Alessandro Verri, Lorenzo Rosasco
We deal with the problem of variable selection when variables must be selected group-wise, with possibly overlapping groups defined a priori. In particular we propose a new optimization procedure for solving the regularized algorithm presented in Jacob et al. 09, where the group lasso penalty is generalized to overlapping groups of variables. While in Jacob et al. 09 the proposed implementation requires explicit replication of the variables belonging to more than one group, our iterative procedure is based on a combination of proximal methods in the primal space and constrained Newton method in a reduced dual space, corresponding to the active groups. This procedure provides a scalable alternative with no need for data duplication, and allows to deal with high dimensional problems without pre-processing to reduce the dimensionality of the data. The computational advantages of our scheme with respect to state-of-the-art algorithms using data duplication are shown empirically with numerical simulations.
****

Heavy-Tailed Process Priors for Selective Shrinkage
Fabian L. Wauthier, Michael Jordan
Heavy-tailed distributions are often used to enhance the robustness of regression and classification methods to outliers in output space. Often, however, we are confronted with ``outliers'' in input space, which are isolated observations in sparsely populated regions. We show that heavy-tailed process priors (which we construct from Gaussian processes via a copula), can be used to improve robustness of regression and classification estimators to such outliers by selectively shrinking them more strongly in sparse regions than in dense regions. We carry out a theoretical analysis to show that selective shrinkage occurs provided the marginals of the heavy-tailed process have sufficiently heavy tails. The analysis is complemented by experiments on biological data which indicate significant improvements of estimates in sparse regions while producing competitive results in dense regions.
****

A New Probabilistic Model for Rank Aggregation
Tao Qin, Xiubo Geng, Tie-yan Liu
This paper is concerned with rank aggregation, which aims to combine multiple input rankings to get a better ranking. A popular approach to rank aggregation is based on probabilistic models on permutations, e.g., the Luce model and the Mallows model. However, these models have their limitations in either poor expressiveness or high computational complexity. To avoid these limitations, in this paper, we propose a new model, which is defined with a coset-permutation distance, and models the generation of a permutation as a stagewise process. We refer to th

e new model as coset-permutation distance based stagewise (CPS) model. The CPS m odel has rich expressiveness and can therefore be used in versatile applications , because many different permutation distances can be used to induce the coset-p ermutation distance. The complexity of the CPS model is low because of the stage wise decomposition of the permutation probability and the efficient computation of most coset-permutation distances. We apply the CPS model to supervised rank a ggregation, derive the learning and inference algorithms, and empirically study their effectiveness and efficiency. Experiments on public datasets show that the derived algorithms based on the CPS model can achieve state-of-the-art ranking accuracy, and are much more efficient than previous algorithms.
****

Learning Networks of Stochastic Differential Equations
José Pereira, Morteza Ibrahimi, Andrea Montanari
We consider linear models for stochastic dynamics. Any such model can be associa ted a network (namely a directed graph) describing which degrees of freedom inte ract under the dynamics. We tackle the problem of learning such a network from o bservation of the system trajectory over a time interval T. We analyse the l1-re gularized least squares algorithm and, in the setting in which the underlying ne twork is sparse, we prove performance guarantees that are uniform in the samplin g rate as long as this is sufficiently high. This result substantiates the notio n of a well defined ‘time complexity’ for the network inference problem.
****

Predictive Subspace Learning for Multi-view Data: a Large Margin Approach
Ning Chen, Jun Zhu, Eric Xing
Learning from multi-view data is important in many applications, such as image c lassification and annotation. In this paper, we present a large-margin learning framework to discover a predictive latent subspace representation shared by mult iple views. Our approach is based on an undirected latent space Markov network t hat fulfills a weak conditional independence assumption that multi-view observat ions and response variables are independent given a set of latent variables. We provide efficient inference and parameter estimation methods for the latent subs pace model. Finally, we demonstrate the advantages of large-margin learning on r eal video and web image data for discovering predictive latent representations a nd improving the performance on image classification, annotation and retrieval.
****

A Bayesian Approach to Concept Drift
Stephen Bach, Mark Maloof
To cope with concept drift, we placed a probability distribution over the locati on of the most-recent drift point. We used Bayesian model comparison to update t his distribution from the predictions of models trained on blocks of consecutive observations and pruned potential drift points with low probability. We compare our approach to a non-probabilistic method for drift and a probabilistic method for change-point detection. In our experiments, our approach generally yielded improved accuracy and/or speed over these other methods.
****

Multivariate Dyadic Regression Trees for Sparse Learning Problems
Han Liu, Xi Chen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

A novel family of non-parametric cumulative based divergences for point processe s
Sohan Seth, Park Il, Austin Brockmeier, Mulugeta Semework, John Choi, Joseph Fra ncis, Jose Principe
Hypothesis testing on point processes has several applications such as model fit ting, plasticity detection, and non-stationarity detection. Standard tools for h ypothesis testing include tests on mean firing rate and time varying rate functi on. However, these statistics do not fully describe a point process and thus the

tests can be misleading. In this paper, we introduce a family of non-parametric divergence measures for hypothesis testing. We extend the traditional Kolmogorov--Smirnov and Cramer--von-Mises tests for point process via stratification. The proposed divergence measures compare the underlying probability structure and, thus, is zero if and only if the point processes are the same. This leads to a more robust test of hypothesis. We prove consistency and show that these measures can be efficiently estimated from data. We demonstrate an application of using the proposed divergence as a cost function to find optimally matched spike trains.

****

Sidestepping Intractable Inference with Structured Ensemble Cascades

David Weiss, Benjamin Sapp, Ben Taskar

For many structured prediction problems, complex models often require adopting approximate inference techniques such as variational methods or sampling, which generally provide no satisfactory accuracy guarantees. In this work, we propose sidestepping intractable inference altogether by learning ensembles of tractable sub-models as part of a structured prediction cascade. We focus in particular on problems with high-treewidth and large state-spaces, which occur in many computer vision tasks. Unlike other variational methods, our ensembles do not enforce agreement between sub-models, but filter the space of possible outputs by simply adding and thresholding the max-marginals of each constituent model. Our framework jointly estimates parameters for all models in the ensemble for each level of the cascade by minimizing a novel, convex loss function, yet requires only a linear increase in computation over learning or inference in a single tractable sub-model. We provide a generalization bound on the filtering loss of the ensemble as a theoretical justification of our approach, and we evaluate our method on both synthetic data and the task of estimating articulated human pose from challenging videos. We find that our approach significantly outperforms loopy belief propagation on the synthetic data and a state-of-the-art model on the pose estimation/tracking problem.

****

Cross Species Expression Analysis using a Dirichlet Process Mixture Model with Latent Matchings

Ziv Bar-joseph, Hai-son Le

Recent studies compare gene expression data across species to identify core and species specific genes in biological systems. To perform such comparisons researchers need to match genes across species. This is a challenging task since the correct matches (orthologs) are not known for most genes. Previous work in this area used deterministic matchings or reduced multidimensional expression data to binary representation. Here we develop a new method that can utilize soft matches (given as priors) to infer both, unique and similar expression patterns across species and a matching for the genes in both species. Our method uses a Dirichlet process mixture model which includes a latent data matching variable. We present learning and inference algorithms based on variational methods for this model. Applying our method to immune response data we show that it can accurately identify common and unique response patterns by improving the matchings between human and mouse genes.

****

t-logistic regression

Nan Ding, S.v.n. Vishwanathan

We extend logistic regression by using t-exponential families which were introduced recently in statistical physics. This gives rise to a regularized risk minimization problem with a non-convex loss function. An efficient block coordinate descent optimization scheme can be derived for estimating the parameters. Because of the nature of the loss function, our algorithm is tolerant to label noise. Furthermore, unlike other algorithms which employ non-convex loss functions, our algorithm is fairly robust to the choice of initial values. We verify both these observations empirically on a number of synthetic and real datasets.

****

Occlusion Detection and Motion Estimation with Convex Optimization

Alper Ayvaci, Michalis Raptis, Stefano Soatto
We tackle the problem of simultaneously detecting occlusions and estimating optical flow. We show that, under standard assumptions of Lambertian reflection and static illumination, the task can be posed as a convex minimization problem. Therefore, the solution, computed using efficient algorithms, is guaranteed to be globally optimal, for any number of independently moving objects, and any number of occlusion layers. We test the proposed algorithm on benchmark datasets, expanded to enable evaluation of occlusion detection performance.
****
Exploiting weakly-labeled Web images to improve object classification: a domain adaptation approach
Alessandro Bergamo, Lorenzo Torresani
Most current image categorization methods require large collections of manually annotated training examples to learn accurate visual recognition models. The time-consuming human labeling effort effectively limits these approaches to recognition problems involving a small number of different object classes. In order to address this shortcoming, in recent years several authors have proposed to learn object classifiers from weakly-labeled Internet images, such as photos retrieved by keyword-based image search engines. While this strategy eliminates the need for human supervision, the recognition accuracies of these methods are considerably lower than those obtained with fully-supervised approaches, because of the noisy nature of the labels associated to Web data. In this paper we investigate and compare methods that learn image classifiers by combining very few manually annotated examples (e.g., 1-10 images per class) and a large number of weakly-labeled Web photos retrieved using keyword-based image search. We cast this as a domain adaptation problem: given a few strongly-labeled examples in a target domain (the manually annotated examples) and many source domain examples (the weakly-labeled Web photos), learn classifiers yielding small generalization error on the target domain. Our experiments demonstrate that, for the same number of strongly-labeled examples, our domain adaptation approach produces significant recognition rate improvements over the best published results (e.g., 65% better when using 5 labeled training examples per class) and that our classifiers are one order of magnitude faster to learn and to evaluate than the best competing method, despite our use of large weakly-labeled data sets.
****
Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression
Ling Huang, Jinzhu Jia, Bin Yu, Byung-gon Chun, Petros Maniatis, Mayur Naik
Predicting the execution time of computer programs is an important but challenging problem in the community of computer systems. Existing methods require experts to perform detailed analysis of program code in order to construct predictors or select important features. We recently developed a new system to automatically extract a large number of features from program execution on sample inputs, on which prediction models can be constructed without expert knowledge. In this paper we study the construction of predictive models for this problem. We propose the SPORE (Sparse POlynomial REgression) methodology to build accurate prediction models of program performance using feature data collected from program execution on sample inputs. Our two SPORE algorithms are able to build relationships between responses (e.g., the execution time of a computer program) and features, and select a few from hundreds of the retrieved features to construct an explicitly sparse and non-linear model to predict the response variable. The compact and explicitly polynomial form of the estimated model could reveal important insights into the computer program (e.g., features and their non-linear combinations that dominate the execution time), enabling a better understanding of the program’s behavior. Our evaluation on three widely used computer programs shows that SPORE methods can give accurate prediction with relative error less than 7% by using a moderate number of training data samples. In addition, we compare SPORE algorithms to state-of-the-art sparse regression algorithms, and show that SPORE methods, motivated by real applications, outperform the other methods in terms of both interpretability and prediction accuracy.

****
Humans Learn Using Manifolds, Reluctantly
Tim Rogers, Chuck Kalish, Joseph Harrison, Jerry Zhu, Bryan Gibson
When the distribution of unlabeled data in feature space lies along a manifold, the information it provides may be used by a learner to assist classification in a semi-supervised setting. While manifold learning is well-known in machine learning, the use of manifolds in human learning is largely unstudied. We perform a set of experiments which test a human's ability to use a manifold in a semi-supervised learning task, under varying conditions. We show that humans may be encouraged into using the manifold, overcoming the strong preference for a simple, axis-parallel linear boundary.
****
Sphere Embedding: An Application to Part-of-Speech Induction
Yariv Maron, Michael Lamar, Elie Bienenstock
Motivated by an application to unsupervised part-of-speech tagging, we present an algorithm for the Euclidean embedding of large sets of categorical data based on co-occurrence statistics. We use the CODE model of Globerson et al. but constrain the embedding to lie on a high-dimensional unit sphere. This constraint allows for efficient optimization, even in the case of large datasets and high embedding dimensionality. Using k-means clustering of the embedded data, our approach efficiently produces state-of-the-art results. We analyze the reasons why the sphere constraint is beneficial in this application, and conjecture that these reasons might apply quite generally to other large-scale tasks.
****
Tight Sample Complexity of Large-Margin Learning
Sivan Sabato, Nathan Srebro, Naftali Tishby
We obtain a tight distribution-specific characterization of the sample complexity of large-margin classification with L2 regularization: We introduce the gamma-adapted-dimension, which is a simple function of the spectrum of a distribution's covariance matrix, and show distribution-specific upper and lower bounds on the sample complexity, both governed by the gamma-adapted-dimension of the source distribution. We conclude that this new quantity tightly characterizes the true sample complexity of large-margin classification. The bounds hold for a rich family of sub-Gaussian distributions.
****
Minimum Average Cost Clustering
Kiyohito Nagano, Yoshinobu Kawahara, Satoru Iwata
A number of objective functions in clustering problems can be described with submodular functions. In this paper, we introduce the minimum average cost criterion, and show that the theory of intersecting submodular functions can be used for clustering with submodular objective functions. The proposed algorithm does not require the number of clusters in advance, and it will be determined by the property of a given set of data points. The minimum average cost clustering problem is parameterized with a real variable, and surprisingly, we show that all information about optimal clusterings for all parameters can be computed in polynomial time in total. Additionally, we evaluate the performance of the proposed algorithm through computational experiments.
****
Online Classification with Specificity Constraints
Andrey Bernstein, Shie Mannor, Nahum Shimkin
We consider the online binary classification problem, where we are given m classifiers. At each stage, the classifiers map the input to the probability that the input belongs to the positive class. An online classification meta-algorithm is an algorithm that combines the outputs of the classifiers in order to attain a certain goal, without having prior knowledge on the form and statistics of the input, and without prior knowledge on the performance of the given classifiers. In this paper, we use sensitivity and specificity as the performance metrics of the meta-algorithm. In particular, our goal is to design an algorithm which satisfies the following two properties (asymptotically): (i) its average false positive rate (fp-rate) is under some given threshold, and (ii) its average true posi

tive rate (tp-rate) is not worse than the tp-rate of the best convex combination of the m given classifiers that satisfies fp-rate constraint, in hindsight. We show that this problem is in fact a special case of the regret minimization problem with constraints, and therefore the above goal is not attainable. Hence, we pose a relaxed goal and propose a corresponding practical online learning meta-algorithm that attains it. In the case of two classifiers, we show that this algorithm takes a very simple form. To our best knowledge, this is the first algorithm that addresses the problem of the average tp-rate maximization under average fp-rate constraints in the online setting.
****

Construction of Dependent Dirichlet Processes based on Poisson Processes
Dahua Lin, Eric Grimson, John Fisher
We present a novel method for constructing dependent Dirichlet processes. The approach exploits the intrinsic relationship between Dirichlet and Poisson pro- cesses in order to create a Markov chain of Dirichlet processes suitable for use as a prior over evolving mixture models. The method allows for the creation, removal, and location variation of component models over time while maintaining the property that the random measures are marginally DP distributed. Additionally, we derive a Gibbs sampling algorithm for model inference and test it on both synthetic and real data. Empirical results demonstrate that the approach is effec- tive in estimating dynamically varying mixture models.
****

Practical Large-Scale Optimization for Max-norm Regularization
Jason D. Lee, Ben Recht, Nathan Srebro, Joel Tropp, Russ R. Salakhutdinov
The max-norm was proposed as a convex matrix regularizer by Srebro et al (2004) and was shown to be empirically superior to the trace-norm for collaborative filtering problems. Although the max-norm can be computed in polynomial time, there are currently no practical algorithms for solving large-scale optimization problems that incorporate the max-norm. The present work uses a factorization technique of Burer and Monteiro (2003) to devise scalable first-order algorithms for convex programs involving the max-norm. These algorithms are applied to solve huge collaborative filtering, graph cut, and clustering problems. Empirically, the new methods outperform mature techniques from all three areas.
****

Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication
Guy Isely, Christopher Hillar, Fritz Sommer
A new algorithm is proposed for a) unsupervised learning of sparse representations from subsampled measurements and b) estimating the parameters required for linearly reconstructing signals from the sparse codes. We verify that the new algorithm performs efficient data compression on par with the recent method of compressive sampling. Further, we demonstrate that the algorithm performs robustly when stacked in several stages or when applied in undercomplete or overcomplete situations. The new algorithm can explain how neural populations in the brain that receive subsampled input through fiber bottlenecks are able to form coherent response properties.
****

Learning Convolutional Feature Hierarchies for Visual Recognition
Koray Kavukcuoglu, Pierre Sermanet, Y-lan Boureau, Karol Gregor, Michael Mathieu, Yann Cun
We propose an unsupervised method for learning multi-stage hierarchies of sparse convolutional features. While sparse coding has become an increasingly popular method for learning visual features, it is most often trained at the patch level. Applying the resulting filters convolutionally results in highly redundant codes because overlapping patches are encoded in isolation. By training convolutionally over large image windows, our method reduces the redudancy between feature vectors at neighboring locations and improves the efficiency of the overall representation. In addition to a linear decoder that reconstructs the image from sparse features, our method trains an efficient feed-forward encoder that predicts quasi-sparse features from the input. While patch-based

training rarely produces anything but oriented edge detectors, we show that convolutional training produces highly diverse filters, including center-surround filters, corner detectors, cross detectors, and oriented grating detectors. We show that using these filters in multi-stage convolutional network architecture improves performance on a number of visual recognition and detection tasks.

****

Multiple Kernel Learning and the SMO Algorithm

Zhaonan Sun, Nawanol Ampornpunt, Manik Varma, S.v.n. Vishwanathan

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

****

Segmentation as Maximum-Weight Independent Set

William Brendel, Sinisa Todorovic

Given an ensemble of distinct, low-level segmentations of an image, our goal is to identify visually meaningful" segments in the ensemble. Knowledge about any specific objects and surfaces present in the image is not available. The selection of image regions occupied by objects is formalized as the maximum-weight independent set (MWIS) problem. MWIS is the heaviest subset of mutually non-adjacent nodes of an attributed graph. We construct such a graph from all segments in the ensemble. Then, MWIS selects maximally distinctive segments that together partition the image. A new MWIS algorithm is presented. The algorithm seeks a solution directly in the discrete domain, instead of relaxing MWIS to a continuous problem, as common in previous work. It iteratively finds a candidate discrete solution of the Taylor series expansion of the original MWIS objective function around the previous solution. The algorithm is shown to converge to a maximum. Our empirical evaluation on the benchmark Berkeley segmentation dataset shows that the new algorithm eliminates the need for hand-picking optimal input parameters of the state-of-the-art segmenters, and outperforms their best, manually optimized results."

****

Static Analysis of Binary Executables Using Structural SVMs

Nikos Karampatziakis

We cast the problem of identifying basic blocks of code in a binary executable as learning a mapping from a byte sequence to a segmentation of the sequence. In general, inference in segmentation models, such as semi-CRFs, can be cubic in the length of the sequence. By taking advantage of the structure of our problem, we derive a linear-time inference algorithm which makes our approach practical, given that even small programs are tens or hundreds of thousands bytes long. Furthermore, we introduce two loss functions which are appropriate for our problem and show how to use structural SVMs to optimize the learned mapping for these losses. Finally, we present experimental results that demonstrate the advantages of our method against a strong baseline.

****

Factorized Latent Spaces with Structured Sparsity

Yangqing Jia, Mathieu Salzmann, Trevor Darrell

Recent approaches to multi-view learning have shown that factorizing the information into parts that are shared across all views and parts that are private to each view could effectively account for the dependencies and independencies between the different input modalities. Unfortunately, these approaches involve minimizing non-convex objective functions. In this paper, we propose an approach to learning such factorized representations inspired by sparse coding techniques. In particular, we show that structured sparsity allows us to address the multi-view learning problem by alternately solving two convex optimization problems. Furthermore, the resulting factorized latent spaces generalize over existing approaches in that they allow :having latent dimensions shared between any subset of the views instead of between all the views only. We show that our approach outperforms state-of-the-art methods on the task of human pose estimation.

****
A unified model of short-range and long-range motion perception
Shuang Wu, Xuming He, Hongjing Lu, Alan L. Yuille
The human vision system is able to effortlessly perceive both short-range and lo ng-range motion patterns in complex dynamic scenes. Previous work has assumed th at two different mechanisms are involved in processing these two types of motion . In this paper, we propose a hierarchical model as a unified framework for mode ling both short-range and long-range motion perception. Our model consists of tw o key components: a data likelihood that proposes multiple motion hypotheses usi ng nonlinear matching, and a hierarchical prior that imposes slowness and spatia l smoothness constraints on the motion field at multiple scales. We tested our m odel on two types of stimuli, random dot kinematograms and multiple-aperture sti muli, both commonly used in human vision research. We demonstrate that the hiera rchical model adequately accounts for human performance in psychophysical experi ments.
****
Structural epitome: a way to summarize one’s visual experience
Nebojsa Jojic, Alessandro Perina, Vittorio Murino
In order to study the properties of total visual input in humans, a single subje ct wore a camera for two weeks capturing, on average, an image every 20 seconds (www.research.microsoft.com/~jojic/aihs). The resulting new dataset contains a m ix of indoor and outdoor scenes as well as numerous foreground objects. Our firs t analysis goal is to create a visual summary of the subject’s two weeks of life using unsupervised algorithms that would automatically discover recurrent scene s, familiar faces or common actions. Direct application of existing algorithms, such as panoramic stitching (e.g. Photosynth) or appearance-based clustering mo dels (e.g. the epitome), is impractical due to either the large dataset size or the dramatic variation in the lighting conditions. As a remedy to these probl ems, we introduce a novel image representation, the “stel epitome,” and an assoc iated efficient learning algorithm. In our model, each image or image patch is c haracterized by a hidden mapping T, which, as in previous epitome models, define s a mapping between the image-coordinates and the coordinates in the large all-I -have-seen" epitome matrix. The limited epitome real-estate forces the mappings of different images to overlap, with this overlap indicating image similarity. H owever, in our model the image similarity does not depend on direct pixel-to-pix el intensity/color/feature comparisons as in previous epitome models, but on spa tial configuration of scene or object parts, as the model is based on the palett e-invariant stel models. As a result, stel epitomes capture structure that is in variant to non-structural changes, such as illumination, that tend to uniformly affect pixels belonging to a single scene or object part."
****
Tree-Structured Stick Breaking for Hierarchical Data
Zoubin Ghahramani, Michael Jordan, Ryan P. Adams
Many data are naturally modeled by an unobserved hierarchical structure. In this paper we propose a flexible nonparametric prior over unknown data hierarchies. The approach uses nested stick-breaking processes to allow for trees of unbounde d width and depth, where data can live at any node and are infinitely exchangeab le. One can view our model as providing infinite mixtures where the components h ave a dependency structure corresponding to an evolutionary diffusion down a tre e. By using a stick-breaking approach, we can apply Markov chain Monte Carlo met hods based on slice sampling to perform Bayesian inference and simulate from the posterior distribution on trees. We apply our method to hierarchical clustering of images and topic modeling of text data.
****
LSTD with Random Projections
Mohammad Ghavamzadeh, Alessandro Lazaric, Odalric Maillard, Rémi Munos
We consider the problem of reinforcement learning in high-dimensional spaces whe n the number of features is bigger than the number of samples. In particular, we study the least-squares temporal difference (LSTD) learning algorithm when a sp ace of low dimension is generated with a random projection from a high-dimension

al space. We provide a thorough theoretical analysis of the LSTD with random pro jections and derive performance bounds for the resulting algorithm. We also show how the error of LSTD with random projections is propagated through the iterati ons of a policy iteration algorithm and provide a performance bound for the resu lting least-squares policy iteration (LSPI) algorithm.
****

Feature Construction for Inverse Reinforcement Learning
Sergey Levine, Zoran Popovic, Vladlen Koltun
The goal of inverse reinforcement learning is to find a reward function for a Ma rkov decision process, given example traces from its optimal policy. Current IRL techniques generally rely on user-supplied features that form a concise basis f or the reward. We present an algorithm that instead constructs reward features f rom a large collection of component features, by building logical conjunctions o f those component features that are relevant to the example policy. Given exampl e traces, the algorithm returns a reward function as well as the constructed fea tures. The reward function can be used to recover a full, deterministic, station ary policy, and the features can be used to transplant the reward function into any novel environment on which the component features are well defined.
****

An analysis on negative curvature induced by singularity in multi-layer neural-n etwork learning
Eiji Mizutani, Stuart Dreyfus
In the neural-network parameter space, an attractive field is likely to be indu ced by singularities. In such a singularity region, first-order gradient learnin g typically causes a long plateau with very little change in the objective func tion value E (hence, a flat region). Therefore, it may be confused with ``attrac tive'' local minima. Our analysis shows that the Hessian matrix of E tends to be indefinite in the vicinity of (perturbed) singular points, suggesting a promisi ng strategy that exploits negative curvature so as to escape from the singularit y plateaus. For numerical evidence, we limit the scope to small examples (some o f which are found in journal papers) that allow us to confirm singularities and the eigenvalues of the Hessian matrix, and for which computation using a descen t direction of negative curvature encounters no plateau. Even for those small p roblems, no efficient methods have been previously developed that avoided platea us.
****

Joint Cascade Optimization Using A Product Of Boosted Classifiers
Leonidas Lefakis, Francois Fleuret
The standard strategy for efficient object detection consists of building a casc ade composed of several binary classifiers. The detection process takes the form of a lazy evaluation of the conjunction of the responses of these classifiers, and concentrates the computation on difficult parts of the image which can not b e trivially rejected. We introduce a novel algorithm to construct jointly the c lassifiers of such a cascade. We interpret the response of a classifier as a pro bability of a positive prediction, and the overall response of the cascade as th e probability that all the predictions are positive. From this noisy-AND model, we derive a consistent loss and a Boosting procedure to optimize that global pro bability on the training set. Such a joint learning allows the individual predi ctors to focus on a more restricted modeling problem, and improves the performan ce compared to a standard cascade. We demonstrate the efficiency of this approac h on face and pedestrian detection with standard data-sets and comparisons with reference baselines.
****

Parallelized Stochastic Gradient Descent
Martin Zinkevich, Markus Weimer, Lihong Li, Alex Smola
With the increase in available data parallel machine learning has become an in creasingly pressing problem. In this paper we present the first parallel stoch astic gradient descent algorithm including a detailed analysis and experimenta l evidence. Unlike prior work on parallel optimization algorithms our varian t comes with parallel acceleration guarantees and it poses no overly tight lat

ency constraints, which might only be available in the multicore setting. Our analysis introduces a novel proof technique --- contractive mappings to quantify the speed of convergence of parameter distributions to their asymptotic limits. As a side effect this answers the question of how quickly stochastic gradient descent algorithms reach the asymptotically normal regime.
****

Shadow Dirichlet for Restricted Probability Modeling
Bela Frigyik, Maya Gupta, Yihua Chen
Although the Dirichlet distribution is widely used, the independence structure of its components limits its accuracy as a model. The proposed shadow Dirichlet distribution manipulates the support in order to model probability mass functions (pmfs) with dependencies or constraints that often arise in real world problems, such as regularized pmfs, monotonic pmfs, and pmfs with bounded variation. We describe some properties of this new class of distributions, provide maximum entropy constructions, give an expectation-maximization method for estimating the mean parameter, and illustrate with real data.
****

MAP estimation in Binary MRFs via Bipartite Multi-cuts
Sashank J. Reddi, Sunita Sarawagi, Sundar Vishwanathan
We propose a new LP relaxation for obtaining the MAP assignment of a binary MRF with pairwise potentials. Our relaxation is derived from reducing the MAP assignment problem to an instance of a recently proposed Bipartite Multi-cut problem where the LP relaxation is guaranteed to provide an O(log k) approximation where k is the number of vertices adjacent to non-submodular edges in the MRF. We then propose a combinatorial algorithm to efficiently solve the LP and also provide a lower bound by concurrently solving its dual to within an approximation. The algorithm is up to an order of magnitude faster and provides better MAP scores and bounds than the state of the art message passing algorithm of [1] that tightens the local marginal polytope with third-order marginal constraints.
****

Approximate Inference by Compilation to Arithmetic Circuits
Daniel Lowd, Pedro Domingos
Arithmetic circuits (ACs) exploit context-specific independence and determinism to allow exact inference even in networks with high treewidth. In this paper, we introduce the first ever approximate inference methods using ACs, for domains where exact inference remains intractable. We propose and evaluate a variety of techniques based on exact compilation, forward sampling, AC structure learning, Markov network parameter learning, variational inference, and Gibbs sampling. In experiments on eight challenging real-world domains, we find that the methods based on sampling and learning work best: one such method (AC2-F) is faster and usually more accurate than loopy belief propagation, mean field, and Gibbs sampling; another (AC2-G) has a running time similar to Gibbs sampling but is consistently more accurate than all baselines.
****

Random Walk Approach to Regret Minimization
Hariharan Narayanan, Alexander Rakhlin
We propose a computationally efficient random walk on a convex body which rapidly mixes to a time-varying Gibbs distribution. In the setting of online convex optimization and repeated games, the algorithm yields low regret and presents a novel efficient method for implementing mixture forecasting strategies.
****

Unsupervised Kernel Dimension Reduction
Meihong Wang, Fei Sha, Michael Jordan
We apply the framework of kernel dimension reduction, originally designed for supervised problems, to unsupervised dimensionality reduction. In this framework, kernel-based measures of independence are used to derive low-dimensional representations that maximally capture information in covariates in order to predict responses. We extend this idea and develop similarly motivated measures for unsupervised problems where covariates and responses are the same. Our empirical studies show that the resulting compact representation yields meaningful and appealin

g visualization and clustering of data. Furthermore, when used in conjunction wi
th supervised learners for classification, our methods lead to lower classificat
ion errors than state-of-the-art methods, especially when embedding data in spac
es of very few dimensions.
****

Multitask Learning without Label Correspondences
Novi Quadrianto, James Petterson, Tibério Caetano, Alex Smola, S.v.n. Vishwanath
an
We propose an algorithm to perform multitask learning where each task has potent
ially distinct label sets and label correspondences are not readily available. T
his is in contrast with existing methods which either assume that the label sets
 shared by different tasks are the same or that there exists a label mapping ora
cle. Our method directly maximizes the mutual information among the labels, and
we show that the resulting objective function can be efficiently optimized using
 existing algorithms. Our proposed approach has a direct application for data in
tegration with different label spaces for the purpose of classification, such as
 integrating Yahoo! and DMOZ web directories.
****

CUR from a Sparse Optimization Viewpoint
Jacob Bien, Ya Xu, Michael W. Mahoney
The CUR decomposition provides an approximation of a matrix X that has low recon
struction error and that is sparse in the sense that the resulting approximation
 lies in the span of only a few columns of X. In this regard, it appears to be s
imilar to many sparse PCA methods. However, CUR takes a randomized algorithmic a
pproach whereas most sparse PCA methods are framed as convex optimization proble
ms. In this paper, we try to understand CUR from a sparse optimization viewpoint
. In particular, we show that CUR is implicitly optimizing a sparse regression o
bjective and, furthermore, cannot be directly cast as a sparse PCA method. We ob
serve that the sparsity attained by CUR possesses an interesting structure, whic
h leads us to formulate a sparse PCA method that achieves a CUR-like sparsity.
****

Robust Clustering as Ensembles of Affinity Relations
Hairong Liu, Longin Latecki, Shuicheng Yan
In this paper, we regard clustering as ensembles of k-ary affinity relations and
 clusters correspond to subsets of objects with maximal average affinity relatio
ns. The average affinity relation of a cluster is relaxed and well approximated
by a constrained homogenous function. We present an efficient procedure to solve
 this optimization problem, and show that the underlying clusters can be robustl
y revealed by using priors systematically constructed from the data. Our method
can automatically select some points to form clusters, leaving other points un-g
rouped; thus it is inherently robust to large numbers of outliers, which has ser
iously limited the applicability of classical methods. Our method also provides
a unified solution to clustering from k-ary affinity relations with $k \geq 2$, that
is, it applies to both graph-based and hypergraph-based clustering problems. Bot
h theoretical analysis and experimental results show the superiority of our meth
od over classical solutions to the clustering problem, especially when there exi
sts a large number of outliers.
****

Optimal learning rates for Kernel Conjugate Gradient regression
Gilles Blanchard, Nicole Krämer
We prove rates of convergence in the statistical sense for kernel-based least sq
uares regression using a conjugate gradient algorithm, where regularization agai
nst overfitting is obtained by early stopping. This method is directly related t
o Kernel Partial Least Squares, a regression method that combines supervised dim
ensionality reduction with least squares projection. The rates depend on two key
 quantities: first, on the regularity of the target regression function and seco
nd, on the effective dimensionality of the data mapped into the kernel space. Lo
wer bounds on attainable rates depending on these two quantities were establishe
d in earlier literature, and we obtain upper bounds for the considered method th
at match these lower bounds (up to a log factor) if the true regression functio

n belongs to the reproducing kernel Hilbert space. If the latter assumption is n ot fulfilled, we obtain similar convergence rates provided additional unlabeled data are available. The order of the learning rates in these two situations matc h state-of-the-art results that were recently obtained for the least squares sup port vector machine and for linear regularization operators.
****

Rates of convergence for the cluster tree
Kamalika Chaudhuri, Sanjoy Dasgupta
For a density f on R^d, a high-density cluster is any connected component of {x: f(x) >= c}, for some c > 0. The set of all high-density clusters form a hierarc hy called the cluster tree of f. We present a procedure for estimating the clust er tree given samples from f. We give finite-sample convergence rates for our al gorithm, as well as lower bounds on the sample complexity of this estimation pro blem.
****

Throttling Poisson Processes
Uwe Dick, Peter Haider, Thomas Vanck, Michael Brückner, Tobias Scheffer
We study a setting in which Poisson processes generate sequences of decision-mak ing events. The optimization goal is allowed to depend on the rate of decision o utcomes; the rate may depend on a potentially long backlog of events and decisio ns. We model the problem as a Poisson process with a throttling policy that enfo rces a data-dependent rate limit and reduce the learning problem to a convex opt imization problem that can be solved efficiently. This problem setting matches a pplications in which damage caused by an attacker grows as a function of the rat e of unsuppressed hostile events. We report on experiments on abuse detection fo r an email service.
****

An Approximate Inference Approach to Temporal Optimization in Optimal Control
Konrad Rawlik, Marc Toussaint, Sethu Vijayakumar
Algorithms based on iterative local approximations present a practical approach to optimal control in robotic systems. However, they generally require the tempo ral parameters (for e.g. the movement duration or the time point of reaching an intermediate goal) to be specified \textit{a priori}. Here, we present a methodo logy that is capable of jointly optimising the temporal parameters in addition t o the control command profiles. The presented approach is based on a Bayesian ca nonical time formulation of the optimal control problem, with the temporal mappi ng from canonical to real time parametrised by an additional control variable. A n approximate EM algorithm is derived that efficiently optimises both the moveme nt duration and control commands offering, for the first time, a practical appro ach to tackling generic via point problems in a systematic way under the optimal control framework. The proposed approach is evaluated on simulations of a redun dant robotic plant.
****

Phone Recognition with the Mean-Covariance Restricted Boltzmann Machine
George Dahl, Marc'aurelio Ranzato, Abdel-rahman Mohamed, Geoffrey E. Hinton
Straightforward application of Deep Belief Nets (DBNs) to acoustic modeling prod uces a rich distributed representation of speech data that is useful for recogni tion and yields impressive results on the speaker-independent TIMIT phone recogn ition task. However, the first-layer Gaussian-Bernoulli Restricted Boltzmann Mac hine (GRBM) has an important limitation, shared with mixtures of diagonal-covari ance Gaussians: GRBMs treat different components of the acoustic input vector as conditionally independent given the hidden state. The mean-covariance restricte d Boltzmann machine (mcRBM), first introduced for modeling natural images, is a much more representationally efficient and powerful way of modeling the covarian ce structure of speech data. Every configuration of the precision units of the m cRBM specifies a different precision matrix for the conditional distribution ove r the acoustic space. In this work, we use the mcRBM to learn features of speech data that serve as input into a standard DBN. The mcRBM features combined with DBNs allow us to achieve a phone error rate of 20.5\%, which is superior to all published results on speaker-independent TIMIT to date.

****
Sparse Coding for Learning Interpretable Spatio-Temporal Primitives
Taehwan Kim, Gregory Shakhnarovich, Raquel Urtasun
Sparse coding has recently become a popular approach in computer vision to learn dictionaries of natural images. In this paper we extend sparse coding to learn interpretable spatio-temporal primitives of human motion. We cast the problem of learning spatio-temporal primitives as a tensor factorization problem and introduce constraints to learn interpretable primitives. In particular, we use group norms over those tensors, diagonal constraints on the activations as well as smoothness constraints that are inherent to human motion. We demonstrate the effectiveness of our approach to learn interpretable representations of human motion from motion capture data, and show that our approach outperforms recently developed matching pursuit and sparse coding algorithms.
****
Block Variable Selection in Multivariate Regression and High-dimensional Causal Inference
Vikas Sindhwani, Aurelie C. Lozano
We consider multivariate regression problems involving high-dimensional predictor and response spaces. To efficiently address such problems, we propose a variable selection method, Multivariate Group Orthogonal Matching Pursuit, which extends the standard Orthogonal Matching Pursuit technique to account for arbitrary sparsity patterns induced by domain-specific groupings over both input and output variables, while also taking advantage of the correlation that may exist between the multiple outputs. We illustrate the utility of this framework for inferring causal relationships over a collection of high-dimensional time series variables. When applied to time-evolving social media content, our models yield a new family of causality-based influence measures that may be seen as an alternative to PageRank. Theoretical guarantees, extensive simulations and empirical studies confirm the generality and value of our framework.
****
Reverse Multi-Label Learning
James Petterson, Tibério Caetano
Multi-label classification is the task of predicting potentially multiple labels for a given instance. This is common in several applications such as image annotation, document classification and gene function prediction. In this paper we present a formulation for this problem based on reverse prediction: we predict sets of instances given the labels. By viewing the problem from this perspective, the most popular quality measures for assessing the performance of multi-label classification admit relaxations that can be efficiently optimised. We optimise these relaxations with standard algorithms and compare our results with several state-of-the-art methods, showing excellent performance.
****
Efficient Relational Learning with Hidden Variable Detection
Ni Lao, Jun Zhu, Liu Liu, Yandong Liu, William W. Cohen
Markov networks (MNs) can incorporate arbitrarily complex features in modeling relational data. However, this flexibility comes at a sharp price of training an exponentially complex model. To address this challenge, we propose a novel relational learning approach, which consists of a restricted class of relational MNs (RMNs) called relation tree-based RMN (treeRMN), and an efficient Hidden Variable Detection algorithm called Contrastive Variable Induction (CVI). On one hand, the restricted treeRMN only considers simple (e.g., unary and pairwise) features in relational data and thus achieves computational efficiency; and on the other hand, the CVI algorithm efficiently detects hidden variables which can capture long range dependencies. Therefore, the resultant approach is highly efficient yet does not sacrifice its expressive power. Empirical results on four real datasets show that the proposed relational learning method can achieve similar prediction quality as the state-of-the-art approaches, but is significantly more efficient in training; and the induced hidden variables are semantically meaningful and crucial to improve the training speed and prediction qualities of treeRMNs.
****

Learning from Logged Implicit Exploration Data
Alex Strehl, John Langford, Lihong Li, Sham M. Kakade
We provide a sound and consistent foundation for the use of \emph{nonrandom} exp loration data in contextual bandit'' orpartially labeled'' settings where only t he value of a chosen action is learned.
The primary challenge in a variety of settings is that the explora tion policy, in which ``offline'' data is logged, is not explicitly known. Prior solutions here require either control of the actions during the learning proces s, recorded random exploration, or actions chosen obliviously in a repeated mann er. The techniques reported here lift these restrictions, allowing the learning of a policy for choosing actions given features from historical data where no r andomization occurred or was logged. We empirically verify our solution on two reasonably sized sets of real-world data obtained from an Internet %online adver tising company.
****
Parametric Bandits: The Generalized Linear Case
Sarah Filippi, Olivier Cappe, Aurélien Garivier, Csaba Szepesvári
We consider structured multi-armed bandit tasks in which the agent is guided by prior structural knowledge that can be exploited to efficiently select the optim al arm(s) in situations where the number of arms is large, or even infinite. We pro- pose a new optimistic, UCB-like, algorithm for non-linearly parameterized b andit problems using the Generalized Linear Model (GLM) framework. We analyze th e regret of the proposed algorithm, termed GLM-UCB, obtaining results similar to those recently proved in the literature for the linear regression case. The ana lysis also highlights a key difficulty of the non-linear case which is solved in GLM-UCB by focusing on the reward space rather than on the parameter space. Mor eover, as the actual efficiency of current parameterized bandit algorithms is of ten deceiving in practice, we provide an asymptotic argument leading to signific antly faster convergence. Simulation studies on real data sets illustrate the pe rformance and the robustness of the proposed GLM-UCB approach.
****
Basis Construction from Power Series Expansions of Value Functions
Sridhar Mahadevan, Bo Liu
This paper explores links between basis construction methods in Markov decision processes and power series expansions of value functions. This perspective provi des a useful framework to analyze properties of existing bases, as well as provi des insight into constructing more effective bases. Krylov and Bellman error bas es are based on the Neumann series expansion. These bases incur very large initi al Bellman errors, and can converge rather slowly as the discount factor approac hes unity. The Laurent series expansion, which relates discounted and average-re ward formulations, provides both an explanation for this slow convergence as wel l as suggests a way to construct more efficient basis representations. The first two terms in the Laurent series represent the scaled average-reward and the ave rage-adjusted sum of rewards, and subsequent terms expand the discounted value f unction using powers of a generalized inverse called the Drazin (or group invers e) of a singular matrix derived from the transition matrix. Experiments show tha t Drazin bases converge considerably more quickly than several other bases, part icularly for large values of the discount factor. An incremental variant of Draz in bases called Bellman average-reward bases (BARBs) is described, which provide s some of the same benefits at lower computational cost.
****
A Novel Kernel for Learning a Neuron Model from Spike Train Data
Nicholas Fisher, Arunava Banerjee
From a functional viewpoint, a spiking neuron is a device that transforms input spike trains on its various synapses into an output spike train on its axon. We demonstrate in this paper that the function mapping underlying the device can be tractably learned based on input and output spike train data alone. We begin by posing the problem in a classification based framework. We then derive a novel kernel for an SRM0 model that is based on PSP and AHP like functions. With the k ernel we demonstrate how the learning problem can be posed as a Quadratic Progra

m. Experimental results demonstrate the strength of our approach.
****

Nonparametric Bayesian Policy Priors for Reinforcement Learning
Finale Doshi-velez, David Wingate, Nicholas Roy, Joshua Tenenbaum
We consider reinforcement learning in partially observable domains where the agent can query an expert for demonstrations. Our nonparametric Bayesian approach combines model knowledge, inferred from expert information and independent exploration, with policy knowledge inferred from expert trajectories. We introduce priors that bias the agent towards models with both simple representations and simple policies, resulting in improved policy and model learning.
****

On the Theory of Learnining with Privileged Information
Dmitry Pechyony, Vladimir Vapnik
In Learning Using Privileged Information (LUPI) paradigm, along with the standard training data in the decision space, a teacher supplies a learner with the privileged information in the correcting space. The goal of the learner is to find a classifier with a low generalization error in the decision space. We consider a new version of empirical risk minimization algorithm, called Privileged ERM, that takes into account the privileged information in order to find a good function in the decision space. We outline the conditions on the correcting space that, if satisfied, allow Privileged ERM to have much faster learning rate in the decision space than the one of the regular empirical risk minimization.
****

Accounting for network effects in neuronal responses using L1 regularized point process models
Ryan Kelly, Matthew Smith, Robert Kass, Tai Lee
Activity of a neuron, even in the early sensory areas, is not simply a function of its local receptive field or tuning properties, but depends on global context of the stimulus, as well as the neural context. This suggests the activity of the surrounding neurons and global brain states can exert considerable influence on the activity of a neuron. In this paper we implemented an L1 regularized point process model to assess the contribution of multiple factors to the firing rate of many individual units recorded simultaneously from V1 with a 96-electrode Utah" array. We found that the spikes of surrounding neurons indeed provide strong predictions of a neuron's response, in addition to the neuron's receptive field transfer function. We also found that the same spikes could be accounted for with the local field potentials, a surrogate measure of global network states. This work shows that accounting for network fluctuations can improve estimates of single trial firing rate and stimulus-response transfer functions."
****

Probabilistic latent variable models for distinguishing between cause and effect
Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M. Mooij, Bernhard Schölkopf
We propose a novel method for inferring whether X causes Y or vice versa from joint observations of X and Y. The basic idea is to model the observed data using probabilistic latent variable models, which incorporate the effects of unobserved noise. To this end, we consider the hypothetical effect variable to be a function of the hypothetical cause variable and an independent noise term (not necessarily additive). An important novel aspect of our work is that we do not restrict the model class, but instead put general non-parametric priors on this function and on the distribution of the cause. The causal direction can then be inferred by using standard Bayesian model selection. We evaluate our approach on synthetic data and real-world data and report encouraging results.
****

Two-Layer Generalization Analysis for Ranking Using Rademacher Average
Wei Chen, Tie-yan Liu, Zhi-ming Ma
This paper is concerned with the generalization analysis on learning to rank for information retrieval (IR). In IR, data are hierarchically organized, i.e., consisting of queries and documents per query. Previous generalization analysis for ranking, however, has not fully considered this structure, and cannot explain how the simultaneous change of query number and document number in the training d

ata will affect the performance of algorithms. In this paper, we propose perform ing generalization analysis under the assumption of two-layer sampling, i.e., th e i.i.d. sampling of queries and the conditional i.i.d sampling of documents per query. Such a sampling can better describe the generation mechanism of real dat a, and the corresponding generalization analysis can better explain the real beh aviors of learning to rank algorithms. However, it is challenging to perform suc h analysis, because the documents associated with different queries are not iden tically distributed, and the documents associated with the same query become no longer independent if represented by features extracted from the matching betwee n document and query. To tackle the challenge, we decompose the generalization e rror according to the two layers, and make use of the new concept of two-layer R ademacher average. The generalization bounds we obtained are quite intuitive and are in accordance with previous empirical studies on the performance of ranking algorithms.

****

Learning from Candidate Labeling Sets

Jie Luo, Francesco Orabona

In many real world applications we do not have access to fully-labeled training data, but only to a list of possible labels. This is the case, e.g., when learni ng visual classifiers from images downloaded from the web, using just their text captions or tags as learning oracles. In general, these problems can be very di fficult. However most of the time there exist different implicit sources of info rmation, coming from the relations between instances and labels, which are usual ly dismissed. In this paper, we propose a semi-supervised framework to model thi s kind of problems. Each training sample is a bag containing multi-instances, as sociated with a set of candidate labeling vectors. Each labeling vector encodes the possible labels for the instances in the bag, with only one being fully corr ect. The use of the labeling vectors provides a principled way not to exclude an y information. We propose a large margin discriminative formulation, and an effi cient algorithm to solve it. Experiments conducted on artificial datasets and a real-world images and captions dataset show that our approach achieves performan ce comparable to SVM trained with the ground-truth labels, and outperforms other baselines.

****

Direct Loss Minimization for Structured Prediction

Tamir Hazan, Joseph Keshet, David McAllester

In discriminative machine learning one is interested in training a system to opt imize a certain desired measure of performance, or loss. In binary classificatio n one typically tries to minimizes the error rate. But in structured prediction each task often has its own measure of performance such as the BLEU score in mac hine translation or the intersection-over-union score in PASCAL segmentation. Th e most common approaches to structured prediction, structural SVMs and CRFs, do not minimize the task loss: the former minimizes a surrogate loss with no guaran tees for task loss and the latter minimizes log loss independent of task loss. T he main contribution of this paper is a theorem stating that a certain perceptro n-like learning rule, involving features vectors derived from loss-adjusted infe rence, directly corresponds to the gradient of task loss. We give empirical resu lts on phonetic alignment of a standard test set from the TIMIT corpus, which su rpasses all previously reported results on this problem.

****

Variational Inference over Combinatorial Spaces

Alexandre Bouchard-côté, Michael Jordan

Since the discovery of sophisticated fully polynomial randomized algorithms for a range of #P problems (Karzanov et al., 1991; Jerrum et al., 2001; Wilson, 2004), theoretical work on approximate inference in combinatorial spaces has focused on Markov chain Monte Carlo methods. Despite their strong theoretical guarante es, the slow running time of many of these randomized algorithms and the restric tive assumptions on the potentials have hindered the applicability of these algo rithms to machine learning. Because of this, in applications to combinatorial s paces simple exact models are often preferred to more complex models that requir

e approximate inference (Siepel et al., 2004). Variational inference would app ear to provide an appealing alternative, given the success of variational method s for graphical models (Wainwright et al., 2008); unfortunately, however, it is not obvious how to develop variational approximations for combinatorial objects such as matchings, partial orders, plane partitions and sequence alignments. W e propose a new framework that extends variational inference to a wide range of combinatorial spaces. Our method is based on a simple assumption: the existence of a tractable measure factorization, which we show holds in many examples. Sim ulations on a range of matching models show that the algorithm is more general a nd empirically faster than a popular fully polynomial randomized algorithm. We also apply the framework to the problem of multiple alignment of protein sequen ces, obtaining state-of-the-art results on the BAliBASE dataset (Thompson et al. , 1999).
****

Latent Variable Models for Predicting File Dependencies in Large-Scale Software Development

Diane Hu, Laurens Maaten, Youngmin Cho, Sorin Lerner, Lawrence Saul

When software developers modify one or more files in a large code base, they mus t also identify and update other related files. Many file dependencies can be de tected by mining the development history of the code base: in essence, groups of related files are revealed by the logs of previous workflows. From data of this form, we show how to detect dependent files by solving a problem in binary matr ix completion. We explore different latent variable models (LVMs) for this probl em, including Bernoulli mixture models, exponential family PCA, restricted Boltz mann machines, and fully Bayesian approaches. We evaluate these models on the de velopment histories of three large, open-source software systems: Mozilla Firefo x, Eclipse Subversive, and Gimp. In all of these applications, we find that LVMs improve the performance of related file prediction over current leading methods .
****

Avoiding False Positive in Multi-Instance Learning

Yanjun Han, Qing Tao, Jue Wang

In multi-instance learning, there are two kinds of prediction failure, i.e., fal se negative and false positive. Current research mainly focus on avoding the for mer. We attempt to utilize the geometric distribution of instances inside positi ve bags to avoid both the former and the latter. Based on kernel principal compo nent analysis, we define a projection constraint for each positive bag to classi fy its constituent instances far away from the separating hyperplane while place positive instances and negative instances at opposite sides. We apply the Const rained Concave-Convex Procedure to solve the resulted problem. Empirical results demonstrate that our approach offers improved generalization performance.
****

Individualized ROI Optimization via Maximization of Group-wise Consistency of St ructural and Functional Profiles

Kaiming Li, Lei Guo, Carlos Faraco, Dajiang Zhu, Fan Deng, Tuo Zhang, Xi Jiang, Degang Zhang, Hanbo Chen, Xintao Hu, Steve Miller, Tianming Liu

Functional segregation and integration are fundamental characteristics of the hu man brain. Studying the connectivity among segregated regions and the dynamics o f integrated brain networks has drawn increasing interest. A very controversial, yet fundamental issue in these studies is how to determine the best functional brain regions or ROIs (regions of interests) for individuals. Essentially, the c omputed connectivity patterns and dynamics of brain networks are very sensitive to the locations, sizes, and shapes of the ROIs. This paper presents a novel met hodology to optimize the locations of an individual's ROIs in the working memory system. Our strategy is to formulate the individual ROI optimization as a group variance minimization problem, in which group-wise functional and structural co nnectivity patterns, and anatomic profiles are defined as optimization constrain ts. The optimization problem is solved via the simulated annealing approach. Our experimental results show that the optimized ROIs have significantly improved c onsistency in structural and functional profiles across subjects, and have more

reasonable localizations and more consistent morphological and anatomic profiles
.
****
On the Convexity of Latent Social Network Inference
Seth Myers, Jure Leskovec
In many real-world scenarios, it is nearly impossible to collect explicit social
 network data. In such cases, whole networks must be inferred from underlying ob
servations. Here, we formulate the problem of inferring latent social networks b
ased on network diffusion or disease propagation data. We consider contagions pr
opagating over the edges of an unobserved social network, where we only observe
the times when nodes became infected, but not who infected them. Given such node
 infection times, we then identify the optimal network that best explains the ob
served data. We present a maximum likelihood approach based on convex programmin
g with a l1-like penalty term that encourages sparsity. Experiments on real and
synthetic data reveal that our method near-perfectly recovers the underlying net
work structure as well as the parameters of the contagion propagation model. Mor
eover, our approach scales well as it can infer optimal networks on thousands of
 nodes in a matter of minutes.
****
Global seismic monitoring as probabilistic inference
Nimar Arora, Stuart J. Russell, Paul Kidwell, Erik Sudderth
The International Monitoring System (IMS) is a global network of sensors whose p
urpose is to identify potential violations of the Comprehensive Nuclear-Test-Ban
 Treaty (CTBT), primarily through detection and localization of seismic events.
We report on the first stage of a project to improve on the current automated so
ftware system with a Bayesian inference system that computes the most likely glo
bal event history given the record of local sensor data. The new system, VISA (V
ertically Integrated Seismological Analysis), is based on empirically calibrated
, generative models of event occurrence, signal propagation, and signal detectio
n. VISA exhibits significantly improved precision and recall compared to the cur
rent operational system and is able to detect events that are missed even by the
 human analysts who post-process the IMS output.
****
Gaussian sampling by local perturbations
George Papandreou, Alan L. Yuille
We present a technique for exact simulation of Gaussian Markov random fields (GM
RFs), which can be interpreted as locally injecting noise to each Gaussian facto
r independently, followed by computing the mean/mode of the perturbed GMRF. Coup
led with standard iterative techniques for the solution of symmetric positive de
finite systems, this yields a very efficient sampling algorithm with essentially
 linear complexity in terms of speed and memory requirements, well suited to ext
remely large scale probabilistic models. Apart from synthesizing data under a Ga
ussian model, the proposed technique directly leads to an efficient unbiased est
imator of marginal variances. Beyond Gaussian models, the proposed algorithm is
also very useful for handling highly non-Gaussian continuously-valued MRFs such
as those arising in statistical image modeling or in the first layer of deep bel
ief networks describing real-valued data, where the non-quadratic potentials cou
pling different sites can be represented as finite or infinite mixtures of Gauss
ians with the help of local or distributed latent mixture assignment variables.
The Bayesian treatment of such models most naturally involves a block Gibbs samp
ler which alternately draws samples of the conditionally independent latent mixt
ure assignments and the conditionally multivariate Gaussian continuous vector an
d we show that it can directly benefit from the proposed methods.
****
Hallucinations in Charles Bonnet Syndrome Induced by Homeostasis: a Deep Boltzma
nn Machine Model
Peggy Series, David Reichert, Amos J. Storkey
The Charles Bonnet Syndrome (CBS) is characterized by complex vivid visual hallu
cinations in people with, primarily, eye diseases and no other neurological path
ology. We present a Deep Boltzmann Machine model of CBS, exploring two core hypo

theses: First, that the visual cortex learns a generative or predictive model of sensory input, thus explaining its capability to generate internal imagery. And second, that homeostatic mechanisms stabilize neuronal activity levels, leading to hallucinations being formed when input is lacking. We reproduce a variety of qualitative findings in CBS. We also introduce a modification to the DBM that allows us to model a possible role of acetylcholine in CBS as mediating the balance of feed-forward and feed-back processing. Our model might provide new insights into CBS and also demonstrates that generative frameworks are promising as hypothetical models of cortical learning and perception.
****

Moreau-Yosida Regularization for Grouped Tree Structure Learning
Jun Liu, Jieping Ye
We consider the tree structured group Lasso where the structure over the features can be represented as a tree with leaf nodes as features and internal nodes as clusters of the features. The structured regularization with a pre-defined tree structure is based on a group-Lasso penalty, where one group is defined for each node in the tree. Such a regularization can help uncover the structured sparsity, which is desirable for applications with some meaningful tree structures on the features. However, the tree structured group Lasso is challenging to solve due to the complex regularization. In this paper, we develop an efficient algorithm for the tree structured group Lasso. One of the key steps in the proposed algorithm is to solve the Moreau-Yosida regularization associated with the grouped tree structure. The main technical contributions of this paper include (1) we show that the associated Moreau-Yosida regularization admits an analytical solution, and (2) we develop an efficient algorithm for determining the effective interval for the regularization parameter. Our experimental results on the AR and JAFFE face data sets demonstrate the efficiency and effectiveness of the proposed algorithm.
****

Pose-Sensitive Embedding by Nonlinear NCA Regression
Graham W. Taylor, Rob Fergus, George Williams, Ian Spiro, Christoph Bregler
This paper tackles the complex problem of visually matching people in similar pose but with different clothes, background, and other appearance changes. We achieve this with a novel method for learning a nonlinear embedding based on several extensions to the Neighborhood Component Analysis (NCA) framework. Our method is convolutional, enabling it to scale to realistically-sized images. By cheaply labeling the head and hands in large video databases through Amazon Mechanical Turk (a crowd-sourcing service), we can use the task of localizing the head and hands as a proxy for determining body pose. We apply our method to challenging real-world data and show that it can generalize beyond hand localization to infer a more general notion of body pose. We evaluate our method quantitatively against other embedding methods. We also demonstrate that real-world performance can be improved through the use of synthetic data.
****

Synergies in learning words and their referents
Mark Johnson, Katherine Demuth, Bevan Jones, Michael Black
This paper presents Bayesian non-parametric models that simultaneously learn to segment words from phoneme strings and learn the referents of some of those words, and shows that there is a synergistic interaction in the acquisition of these two kinds of linguistic information. The models themselves are novel kinds of Adaptor Grammars that are an extension of an embedding of topic models into PCFGs. These models simultaneously segment phoneme sequences into words and learn the relationship between non-linguistic objects to the words that refer to them. We show (i) that modelling inter-word dependencies not only improves the accuracy of the word segmentation but also of word-object relationships, and (ii) that a model that simultaneously learns word-object relationships and word segmentation segments more accurately than one that just learns word segmentation on its own. We argue that these results support an interactive view of language acquisition that can take advantage of synergies such as these.
****

Approximate inference in continuous time Gaussian-Jump processes
Manfred Opper, Andreas Ruttor, Guido Sanguinetti
We present a novel approach to inference in conditionally Gaussian continuous ti me stochastic processes, where the latent process is a Markovian jump process. W e first consider the case of jump-diffusion processes, where the drift of a line ar stochastic differential equation can jump at arbitrary time points. We derive partial differential equations for exact inference and present a very efficient mean field approximation. By introducing a novel lower bound on the free energy , we then generalise our approach to Gaussian processes with arbitrary covarianc e, such as the non-Markovian RBF covariance. We present results on both simulate d and real data, showing that the approach is very accurate in capturing latent dynamics and can be useful in a number of real data modelling tasks.
****
Empirical Bernstein Inequalities for U-Statistics
Thomas Peel, Sandrine Anthoine, Liva Ralaivola
We present original empirical Bernstein inequalities for U-statistics with bound ed symmetric kernels q. They are expressed with respect to empirical estimates o f either the variance of q or the conditional variance that appears in the Berns tein-type inequality for U-statistics derived by Arcones [2]. Our result subsume s other existing empirical Bernstein inequalities, as it reduces to them when U- statistics of order 1 are considered. In addition, it is based on a rather direc t argument using two applications of the same (non-empirical) Bernstein inequali ty for U-statistics. We discuss potential applications of our new inequalities, especially in the realm of learning ranking/scoring functions. In the process, w e exhibit an efficient procedure to compute the variance estimates for the speci al case of bipartite ranking that rests on a sorting argument. We also argue tha t our results may provide test set bounds and particularly interesting empirical racing algorithms for the problem of online learning of scoring functions.
****
Active Estimation of F-Measures
Christoph Sawade, Niels Landwehr, Tobias Scheffer
We address the problem of estimating the F-measure of a given model as accuratel y as possible on a fixed labeling budget. This problem occurs whenever an estima te cannot be obtained from held-out training data; for instance, when data that have been used to train the model are held back for reasons of privacy or do not reflect the test distribution. In this case, new test instances have to be draw n and labeled at a cost. An active estimation procedure selects instances accord ing to an instrumental sampling distribution. An analysis of the sources of esti mation error leads to an optimal sampling distribution that minimizes estimator variance. We explore conditions under which active estimates of F-measures are m ore accurate than estimates based on instances sampled from the test distributio n.
****
Inductive Regularized Learning of Kernel Functions
Prateek Jain, Brian Kulis, Inderjit Dhillon
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Active Learning Applied to Patient-Adaptive Heartbeat Classification
Jenna Wiens, John Guttag
While clinicians can accurately identify different types of heartbeats in electr ocardiograms (ECGs) from different patients, researchers have had limited succes s in applying supervised machine learning to the same task. The problem is made challenging by the variety of tasks, inter- and intra-patient differences, an of ten severe class imbalance, and the high cost of getting cardiologists to label data for individual patients. We address these difficulties using active learnin g to perform patient-adaptive and task-adaptive heartbeat classification. When t ested on a benchmark database of cardiologist annotated ECG recordings, our meth

od had considerably better performance than other recently proposed methods on t
he two primary classification tasks recommended by the Association for the Advan
cement of Medical Instrumentation. Additionally, our method required over 90% le
ss patient-specific training data than the methods to which we compared it.
****

Large-Scale Matrix Factorization with Missing Data under Additional Constraints
Kaushik Mitra, Sameer Sheorey, Rama Chellappa
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Probabilistic Deterministic Infinite Automata
David Pfau, Nicholas Bartlett, Frank Wood
We propose a novel Bayesian nonparametric approach to learning with probabilisti
c deterministic finite automata (PDFA). We define and develop and sampler for a
PDFA with an infinite number of states which we call the probabilistic determini
stic infinite automata (PDIA). Posterior predictive inference in this model, giv
en a finite training sequence, can be interpreted as averaging over multiple PDF
As of varying structure, where each PDFA is biased towards having few states. We
 suggest that our method for averaging over PDFAs is a novel approach to predict
ive distribution smoothing. We test PDIA inference both on PDFA structure learni
ng and on both natural language and DNA data prediction tasks. The results sugge
st that the PDIA presents an attractive compromise between the computational cos
t of hidden Markov models and the storage requirements of hierarchically smoothe
d Markov models.
****

Brain covariance selection: better individual functional connectivity models usi
ng population prior
Gael Varoquaux, Alexandre Gramfort, Jean-baptiste Poline, Bertrand Thirion
Spontaneous brain activity, as observed in functional neuroimaging, has been sho
wn to display reproducible structure that expresses brain architecture and carri
es markers of brain pathologies. An important view of modern neuroscience is tha
t such large-scale structure of coherent activity reflects modularity properties
 of brain connectivity graphs. However, to date, there has been no demonstration
 that the limited and noisy data available in spontaneous activity observations
could be used to learn full-brain probabilistic models that generalize to new da
ta. Learning such models entails two main challenges: i) modeling full brain c
onnectivity is a difficult estimation problem that faces the curse of dimensiona
lity and ii) variability between subjects, coupled with the variability of func
tional signals between experimental runs, makes the use of multiple datasets cha
llenging. We describe subject-level brain functional connectivity structure as
a multivariate Gaussian process and introduce a new strategy to estimate it from
 group data, by imposing a common structure on the graphical model in the popula
tion. We show that individual models learned from functional Magnetic Resonance
Imaging (fMRI) data using this population prior generalize better to unseen data
 than models based on alternative regularization schemes. To our knowledge, this
 is the first report of a cross-validated model of spontaneous brain activity. F
inally, we use the estimated graphical model to explore the large-scale characte
ristics of functional architecture and show for the first time that known cognit
ive networks appear as the integrated communities of functional connectivity gra
ph.
****

Word Features for Latent Dirichlet Allocation
James Petterson, Wray Buntine, Shravan Narayanamurthy, Tibério Caetano, Alex Smo
la
We extend Latent Dirichlet Allocation (LDA) by explicitly allowing for the encod
ing of side information in the distribution over words. This results in a variet
y of new capabilities, such as improved estimates for infrequently occurring wor
ds, as well as the ability to leverage thesauri and dictionaries in order to boo

st topic cohesion within and across languages. We present experiments on multi-l
anguage topic synchronisation where dictionary information is used to bias corre
sponding words towards similar topics. Results indicate that our model substanti
ally improves topic cohesion when compared to the standard LDA model.
****

A Primal-Dual Message-Passing Algorithm for Approximated Large Scale Structured
Prediction
Tamir Hazan, Raquel Urtasun
In this paper we propose an approximated learning framework for large scale grap
hical models and derive message passing algorithms for learning their parameters
 efficiently. We first relate CRFs and structured SVMs and show that in the CR
F's primal a variant of the log-partition function, known as soft-max, smoothly
approximates the hinge loss function of structured SVMs. We then propose an int
uitive approximation for structured prediction problems using Fenchel duality ba
sed on a local entropy approximation that computes the exact gradients of the ap
proximated problem and is guaranteed to converge. Unlike existing approaches, th
is allow us to learn graphical models with cycles and very large number of param
eters efficiently. We demonstrate the effectiveness of our approach in an image
 denoising task. This task was previously solved by sharing parameters across cl
iques. In contrast, our algorithm is able to efficiently learn large number of p
arameters resulting in orders of magnitude better prediction.
****

Efficient algorithms for learning kernels from multiple similarity matrices with
 general convex loss functions
Achintya Kundu, Vikram Tankasali, Chiranjib Bhattacharyya, Aharon Ben-tal
In this paper we consider the problem of learning an n x n Kernel matrix from m
similarity matrices under general convex loss. Past research have extensively st
udied the m =1 case and have derived several algorithms which require sophistica
ted techniques like ACCP, SOCP, etc. The existing algorithms do not apply if one
 uses arbitrary losses and often can not handle m > 1 case. We present several p
rovably convergent iterative algorithms, where each iteration requires either an
 SVM or a Multiple Kernel Learning (MKL) solver for m > 1 case. One of the major
 contributions of the paper is to extend the well known Mirror Descent(MD) frame
work to handle Cartesian product of psd matrices. This novel extension leads to
an algorithm, called EMKL, which solves the problem in O(m^2 log n) iterations;
in each iteration one solves an MKL involving m kernels and m eigen-decompositio
n of n x n matrices. By suitably defining a restriction on the objective functio
n, a faster version of EMKL is proposed, called REKL, which avoids the eigen-dec
omposition. An alternative to both EMKL and REKL is also suggested which require
s only an SVM solver. Experimental results on real world protein data set involv
ing several similarity matrices illustrate the efficacy of the proposed algorith
ms.
****

Online Learning: Random Averages, Combinatorial Parameters, and Learnability
Alexander Rakhlin, Karthik Sridharan, Ambuj Tewari
We develop a theory of online learning by defining several complexity measures.
Among them are analogues of Rademacher complexity, covering numbers and fat-shat
tering dimension from statistical learning theory. Relationship among these comp
lexity measures, their connection to online learning, and tools for bounding the
m are provided. We apply these results to various learning problems. We provide
a complete characterization of online learnability in the supervised setting.
****

Variable margin losses for classifier design
Hamed Masnadi-shirazi, Nuno Vasconcelos
The problem of controlling the margin of a classifier is studied. A detailed ana
lytical study is presented on how properties of the classification risk, such as
 its optimal link and minimum risk functions, are related to the shape of the lo
ss, and its margin enforcing properties. It is shown that for a class of risks,
denoted canonical risks, asymptotic Bayes consistency is compatible with simple
analytical relationships between these functions. These enable a precise charact

erization of the loss for a popular class of link functions. It is shown that, when the risk is in canonical form and the link is inverse sigmoidal, the margin properties of the loss are determined by a single parameter. Novel families of Bayes consistent loss functions, of variable margin, are derived. These families are then used to design boosting style algorithms with explicit control of the classification margin. The new algorithms generalize well established approaches, such as LogitBoost. Experimental results show that the proposed variable margin losses outperform the fixed margin counterparts used by existing algorithms. Finally, it is shown that best performance can be achieved by cross-validating the margin parameter.

****

Nonparametric Density Estimation for Stochastic Optimization with an Observable State Variable

Lauren Hannah, Warren Powell, David Blei

We study convex stochastic optimization problems where a noisy objective function value is observed after a decision is made. There are many stochastic optimization problems whose behavior depends on an exogenous state variable which affects the shape of the objective function. Currently, there is no general purpose algorithm to solve this class of problems. We use nonparametric density estimation for the joint distribution of state-outcome pairs to create weights for previous observations. The weights effectively group similar states. Those similar to the current state are used to create a convex, deterministic approximation of the objective function. We propose two solution methods that depend on the problem characteristics: function-based and gradient-based optimization. We offer two weighting schemes, kernel based weights and Dirichlet process based weights, for use with the solution methods. The weights and solution methods are tested on a synthetic multi-product newsvendor problem and the hour ahead wind commitment problem. Our results show Dirichlet process weights can offer substantial benefits over kernel based weights and, more generally, that nonparametric estimation methods provide good solutions to otherwise intractable problems.

****

Mixture of time-warped trajectory models for movement decoding

Elaine Corbett, Eric Perreault, Konrad Koerding

Applications of Brain-Machine-Interfaces typically estimate user intent based on biological signals that are under voluntary control. For example, we might want to estimate how a patient with a paralyzed arm wants to move based on residual muscle activity. To solve such problems it is necessary to integrate obtained information over time. To do so, state of the art approaches typically use a probabilistic model of how the state, e.g. position and velocity of the arm, evolves over time – a so-called trajectory model. We wanted to further develop this approach using two intuitive insights: (1) At any given point of time there may be a small set of likely movement targets, potentially identified by the location of objects in the workspace or by gaze information from the user. (2) The user may want to produce movements at varying speeds. We thus use a generative model with a trajectory model incorporating these insights. Approximate inference on that generative model is implemented using a mixture of extended Kalman filters. We find that the resulting algorithm allows us to decode arm movements dramatically better than when we use a trajectory model with linear dynamics.

****

A Discriminative Latent Model of Image Region and Object Tag Correspondence

Yang Wang, Greg Mori

We propose a discriminative latent model for annotating images with unaligned object-level textual annotations. Instead of using the bag-of-words image representation currently popular in the computer vision community, our model explicitly captures more intricate relationships underlying visual and textual information. In particular, we model the mapping that translates image regions to annotations. This mapping allows us to relate image regions to their corresponding annotation terms. We also model the overall scene label as latent information. This allows us to cluster test images. Our training data consist of images and their associated annotations. But we do not have access to the ground-truth region-to-ann

otation mapping or the overall scene label. We develop a novel variant of the la tent SVM framework to model them as latent variables. Our experimental results d emonstrate the effectiveness of the proposed model compared with other baseline methods.
****

Lower Bounds on Rate of Convergence of Cutting Plane Methods
Xinhua Zhang, Ankan Saha, S.v.n. Vishwanathan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Self-Paced Learning for Latent Variable Models
M. Kumar, Benjamin Packer, Daphne Koller
Latent variable models are a powerful tool for addressing several tasks in machi ne learning. However, the algorithms for learning the parameters of latent varia ble models are prone to getting stuck in a bad local optimum. To alleviate this problem, we build on the intuition that, rather than considering all samples sim ultaneously, the algorithm should be presented with the training data in a meani ngful order that facilitates learning. The order of the samples is determined by how easy they are. The main challenge is that often we are not provided with a readily computable measure of the easiness of samples. We address this issue by proposing a novel, iterative self-paced learning algorithm where each iteration simultaneously selects easy samples and learns a new parameter vector. The numb er of samples selected is governed by a weight that is annealed until the entire training data has been considered. We empirically demonstrate that the self-pac ed learning algorithm outperforms the state of the art method for learning a lat ent structural SVM on four applications: object localization, noun phrase corefe rence, motif finding and handwritten digit recognition.
****

Learning Efficient Markov Networks
Vibhav Gogate, William Webb, Pedro Domingos
We present an algorithm for learning high-treewidth Markov networks where infere nce is still tractable. This is made possible by exploiting context specific ind ependence and determinism in the domain. The class of models our algorithm can l earn has the same desirable properties as thin junction trees: polynomial infere nce, closed form weight learning, etc., but is much broader. Our algorithm searc hes for a feature that divides the state space into subspaces where the remainin g variables decompose into independent subsets (conditioned on the feature or it s negation) and recurses on each subspace/subset of variables until no useful ne w features can be found. We provide probabilistic performance guarantees for our algorithm under the assumption that the maximum feature length is k (the treewi dth can be much larger) and dependences are of bounded strength. We also propose a greedy version of the algorithm that, while forgoing these guarantees, is muc h more efficient.Experiments on a variety of domains show that our approach comp ares favorably with thin junction trees and other Markov network structure learn ers.
****

Multi-Stage Dantzig Selector
Ji Liu, Peter Wonka, Jieping Ye
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Batch Bayesian Optimization via Simulation Matching
Javad Azimi, Alan Fern, Xiaoli Fern
Bayesian optimization methods are often used to optimize unknown functions that are costly to evaluate. Typically, these methods sequentially select inputs to b e evaluated one at a time based on a posterior over the unknown function that is

updated after each evaluation. There are a number of effective sequential policies for selecting the individual inputs. In many applications, however, it is desirable to perform multiple evaluations in parallel, which requires selecting batches of multiple inputs to evaluate at once. In this paper, we propose a novel approach to batch Bayesian optimization, providing a policy for selecting batches of inputs with the goal of optimizing the function as efficiently as possible. The key idea is to exploit the availability of high-quality and efficient sequential policies, by using Monte-Carlo simulation to select input batches that closely match their expected behavior. To the best of our knowledge, this is the first batch selection policy for Bayesian optimization. Our experimental results on six benchmarks show that the proposed approach significantly outperforms two baselines and can lead to large advantages over a top sequential approach in terms of performance per unit time.

****

Inference with Multivariate Heavy-Tails in Linear Models

Danny Bickson, Carlos Guestrin

Heavy-tailed distributions naturally occur in many real life problems. Unfortunately, it is typically not possible to compute inference in closed-form in graphical models which involve such heavy tailed distributions. In this work, we propose a novel simple linear graphical model for independent latent random variables, called linear characteristic model (LCM), defined in the characteristic function domain. Using stable distributions, a heavy-tailed family of distributions which is a generalization of Cauchy, L\'evy and Gaussian distributions, we show for the first time, how to compute both exact and approximate inference in such a linear multivariate graphical model. LCMs are not limited to only stable distributions, in fact LCMs are always defined for any random variables (discrete, continuous or a mixture of both). We provide a realistic problem from the field of computer networks to demonstrate the applicability of our construction. Other potential application is iterative decoding of linear channels with non-Gaussian noise.

****

Towards Holistic Scene Understanding: Feedback Enabled Cascaded Classification Models

Congcong Li, Adarsh Kowdle, Ashutosh Saxena, Tsuhan Chen

In many machine learning domains (such as scene understanding), several related sub-tasks (such as scene categorization, depth estimation, object detection) operate on the same raw data and provide correlated outputs. Each of these tasks is often notoriously hard, and state-of-the-art classifiers already exist for many sub-tasks. It is desirable to have an algorithm that can capture such correlation without requiring to make any changes to the inner workings of any classifier. We propose Feedback Enabled Cascaded Classification Models (FE-CCM), that maximizes the joint likelihood of the sub-tasks, while requiring only a ‘black-box’ interface to the original classifier for each sub-task. We use a two-layer cascade of classifiers, which are repeated instantiations of the original ones, with the output of the first layer fed into the second layer as input. Our training method involves a feedback step that allows later classifiers to provide earlier classifiers information about what error modes to focus on. We show that our method significantly improves performance in all the sub-tasks in two different domains: (i) scene understanding, where we consider depth estimation, scene categorization, event categorization, object detection, geometric labeling and saliency detection, and (ii) robotic grasping, where we consider grasp point detection and object classification.

****

Learning the context of a category

Dan Navarro

This paper outlines a hierarchical Bayesian model for human category learning that learns both the organization of objects into categories, and the context in which this knowledge should be applied. The model is fit to multiple data sets, and provides a parsimonious method for describing how humans learn context specific conceptual representations.

****
Learning via Gaussian Herding
Koby Crammer, Daniel Lee
We introduce a new family of online learning algorithms based upon constraining the velocity flow over a distribution of weight vectors. In particular, we show how to effectively herd a Gaussian weight vector distribution by trading off velocity constraints with a loss function. By uniformly bounding this loss function, we demonstrate how to solve the resulting optimization analytically. We compare the resulting algorithms on a variety of real world datasets, and demonstrate how these algorithms achieve state-of-the-art robust performance, especially with high label noise in the training data.
****
Monte-Carlo Planning in Large POMDPs
David Silver, Joel Veness
This paper introduces a Monte-Carlo algorithm for online planning in large POMDPs. The algorithm combines a Monte-Carlo update of the agent's belief state with a Monte-Carlo tree search from the current belief state. The new algorithm, POMCP, has two important properties. First, Monte-Carlo sampling is used to break the curse of dimensionality both during belief state updates and during planning. Second, only a black box simulator of the POMDP is required, rather than explicit probability distributions. These properties enable POMCP to plan effectively in significantly larger POMDPs than has previously been possible. We demonstrate its effectiveness in three large POMDPs. We scale up a well-known benchmark problem, Rocksample, by several orders of magnitude. We also introduce two challenging new POMDPs: 10x10 Battleship and Partially Observable PacMan, with approximately 10^18 and 10^56 states respectively. Our Monte-Carlo planning algorithm achieved a high level of performance with no prior knowledge, and was also able to exploit simple domain knowledge to achieve better results with less search. POMCP is the first general purpose planner to achieve high performance in such large and unfactored POMDPs.
****
Spatial and anatomical regularization of SVM for brain image analysis
Remi Cuingnet, Marie Chupin, Habib Benali, Olivier Colliot
Support vector machines (SVM) are increasingly used in brain image analyses since they allow capturing complex multivariate relationships in the data. Moreover, when the kernel is linear, SVMs can be used to localize spatial patterns of discrimination between two groups of subjects. However, the features' spatial distribution is not taken into account. As a consequence, the optimal margin hyperplane is often scattered and lacks spatial coherence, making its anatomical interpretation difficult. This paper introduces a framework to spatially regularize SVM for brain image analysis. We show that Laplacian regularization provides a flexible framework to integrate various types of constraints and can be applied to both cortical surfaces and 3D brain images. The proposed framework is applied to the classification of MR images based on gray matter concentration maps and cortical thickness measures from 30 patients with Alzheimer's disease and 30 elderly controls. The results demonstrate that the proposed method enables natural spatial and anatomical regularization of the classifier.
****
Divisive Normalization: Justification and Effectiveness as Efficient Coding Transform
Siwei Lyu
Divisive normalization (DN) has been advocated as an effective nonlinear {\em efficient coding} transform for natural sensory signals with applications in biology and engineering. In this work, we aim to establish a connection between the DN transform and the statistical properties of natural sensory signals. Our analysis is based on the use of multivariate {\em t} model to capture some important statistical properties of natural sensory signals. The multivariate {\em t} model justifies DN as an approximation to the transform that completely eliminates its statistical dependency. Furthermore, using the multivariate {\em t} model and measuring statistical dependency with multi-information, we can precisely quant

ify the statistical dependency that is reduced by the DN transform. We compare t
his with the actual performance of the DN transform in reducing statistical depe
ndencies of natural sensory signals. Our theoretical analysis and quantitative e
valuations confirm DN as an effective efficient coding transform for natural sen
sory signals. On the other hand, we also observe a previously unreported phenome
non that DN may increase statistical dependencies when the size of pooling is sm
all.
****

Rescaling, thinning or complementing? On goodness-of-fit procedures for point pr
ocess models and Generalized Linear Models
Felipe Gerhard, Wulfram Gerstner
Generalized Linear Models (GLMs) are an increasingly popular framework for model
ing neural spike trains. They have been linked to the theory of stochastic point
 processes and researchers have used this relation to assess goodness-of-fit usi
ng methods from point-process theory, e.g. the time-rescaling theorem. However,
high neural firing rates or coarse discretization lead to a breakdown of the ass
umptions necessary for this connection. Here, we show how goodness-of-fit tests
from point-process theory can still be applied to GLMs by constructing equivalen
t surrogate point processes out of time-series observations. Furthermore, two ad
ditional tests based on thinning and complementing point processes are introduce
d. They augment the instruments available for checking model adequacy of point p
rocesses as well as discretized models.
****

Active Instance Sampling via Matrix Partition
Yuhong Guo
Recently, batch-mode active learning has attracted a lot of attention. In this p
aper, we propose a novel batch-mode active learning approach that selects a batc
h of queries in each iteration by maximizing a natural form of mutual informatio
n criterion between the labeled and unlabeled instances. By employing a Gaussian
 process framework, this mutual information based instance selection problem can
 be formulated as a matrix partition problem. Although the matrix partition is a
n NP-hard combinatorial optimization problem, we show a good local solution can
be obtained by exploiting an effective local optimization technique on the relax
ed continuous optimization problem. The proposed active learning approach is ind
ependent of employed classification models. Our empirical studies show this appr
oach can achieve comparable or superior performance to discriminative batch-mode
 active learning methods.
****

Functional form of motion priors in human motion perception
Hongjing Lu, Tungyou Lin, Alan Lee, Luminita Vese, Alan L. Yuille
It has been speculated that the human motion system combines noisy measurements
with prior expectations in an optimal, or rational, manner. The basic goal of ou
r work is to discover experimentally which prior distribution is used. More spec
ifically, we seek to infer the functional form of the motion prior from the perf
ormance of human subjects on motion estimation tasks. We restricted ourselves to
 priors which combine three terms for motion slowness, first-order smoothness, a
nd second-order smoothness. We focused on two functional forms for prior distrib
utions: L2-norm and L1-norm regularization corresponding to the Gaussian and Lap
lace distributions respectively. In our first experimental session we estimate t
he weights of the three terms for each functional form to maximize the fit to hu
man performance. We then measured human performance for motion tasks and found t
hat we obtained better fit for the L1-norm (Laplace) than for the L2-norm (Gauss
ian). We note that the L1-norm is also a better fit to the statistics of motion
in natural environments. In addition, we found large weights for the second-orde
r smoothness term, indicating the importance of high-order smoothness compared t
o slowness and lower-order smoothness. To validate our results further, we used
the best fit models using the L1-norm to predict human performance in a second s
ession with different experimental setups. Our results showed excellent agreemen
t between human performance and model prediction -- ranging from 3\% to 8\% for
five human subjects over ten experimental conditions -- and give further support

that the human visual system uses an L1-norm (Laplace) prior.
****
The LASSO risk: asymptotic results and real world examples
Mohsen Bayati, José Pereira, Andrea Montanari
We consider the problem of learning a coefficient vector x0 from noisy linear ob servation y=Ax0+w. In many contexts (ranging from model selection to image proce ssing) it is desirable to construct a sparse estimator. In this case, a popular approach consists in solving an l1-penalized least squares problem known as the LASSO or BPDN. For sequences of matrices A of increasing dimensions, with iid g aussian entries, we prove that the normalized risk of the LASSO converges to a l imit, and we obtain an explicit expression for this limit. Our result is the fir st rigorous derivation of an explicit formula for the asymptotic risk of the LAS SO for random instances. The proof technique is based on the analysis of AMP, a recently developed efficient algorithm, that is inspired from graphical models i deas. Through simulations on real data matrices (gene expression data and hosp ital medical records) we observe that these results can be relevant in a broad a rray of practical applications.
****
Layered image motion with explicit occlusions, temporal consistency, and depth o rdering
Deqing Sun, Erik Sudderth, Michael Black
Layered models are a powerful way of describing natural scenes containing smooth surfaces that may overlap and occlude each other. For image motion estimation, such models have a long history but have not achieved the wide use or accuracy o f non-layered methods. We present a new probabilistic model of optical flow in l ayers that addresses many of the shortcomings of previous approaches. In particu lar, we define a probabilistic graphical model that explicitly captures: 1) occl usions and disocclusions; 2) depth ordering of the layers; 3) temporal consisten cy of the layer segmentation. Additionally the optical flow in each layer is mod eled by a combination of a parametric model and a smooth deviation based on an M RF with a robust spatial prior; the resulting model allows roughness in layers. Finally, a key contribution is the formulation of the layers using an image-depe ndent hidden field prior based on recent models for static scene segmentation. T he method achieves state-of-the-art results on the Middlebury benchmark and prod uces meaningful scene segmentations as well as detected occlusion regions.
****
Towards Property-Based Classification of Clustering Paradigms
Margareta Ackerman, Shai Ben-David, David Loker
Clustering is a basic data mining task with a wide variety of applications. Not surprisingly, there exist many clustering algorithms. However, clustering is an ill defined problem - given a data set, it is not clear what a “correct” cluster ing for that set is. Indeed, different algorithms may yield dramatically differe nt outputs for the same input sets. Faced with a concrete clustering task, a use r needs to choose an appropriate clustering algorithm. Currently, such decisions are often made in a very ad hoc, if not completely random, manner. Given the cr ucial effect of the choice of a clustering algorithm on the resulting clustering , this state of affairs is truly regrettable. In this paper we address the major research challenge of developing tools for helping users make more informed dec isions when they come to pick a clustering tool for their data. This is, of cour se, a very ambitious endeavor, and in this paper, we make some first steps towar ds this goal. We propose to address this problem by distilling abstract properti es of the input-output behavior of different clustering paradigms. In this pape r, we demonstrate how abstract, intuitive properties of clustering functions can be used to taxonomize a set of popular clustering algorithmic paradigms. On top of addressing deterministic clustering algorithms, we also propose similar prop erties for randomized algorithms and use them to highlight functional difference s between different common implementations of k-means clustering. We also study relationships between the properties, independent of any particular algorithm. I n particular, we strengthen Kleinbergs famous impossibility result, while provid ing a simpler proof.

****
Implicit encoding of prior probabilities in optimal neural populations
Deep Ganguli, Eero Simoncelli
Optimal coding provides a guiding principle for understanding the representation of sensory variables in neural populations. Here we consider the influence of a prior probability distribution over sensory variables on the optimal allocation of cells and spikes in a neural population. We model the spikes of each cell as samples from an independent Poisson process with rate governed by an associated tuning curve. For this response model, we approximate the Fisher information in terms of the density and amplitude of the tuning curves, under the assumption that tuning width varies inversely with cell density. We consider a family of objective functions based on the expected value, over the sensory prior, of a functional of the Fisher information. This family includes lower bounds on mutual information and perceptual discriminability as special cases. In all cases, we find a closed form expression for the optimum, in which the density and gain of the cells in the population are power law functions of the stimulus prior. This also implies a power law relationship between the prior and perceptual discriminability. We show preliminary evidence that the theory successfully predicts the relationship between empirically measured stimulus priors, physiologically measured neural response properties (cell density, tuning widths, and firing rates), and psychophysically measured discrimination thresholds.
****
Distributed Dual Averaging In Networks
Alekh Agarwal, Martin J. Wainwright, John C. Duchi
The goal of decentralized optimization over a network is to optimize a global objective formed by a sum of local (possibly nonsmooth) convex functions using only local computation and communication. We develop and analyze distributed algorithms based on dual averaging of subgradients, and we provide sharp bounds on their convergence rates as a function of the network size and topology. Our analysis clearly separates the convergence of the optimization algorithm itself from the effects of communication constraints arising from the network structure. We show that the number of iterations required by our algorithm scales inversely in the spectral gap of the network. The sharpness of this prediction is confirmed both by theoretical lower bounds and simulations for various networks.
****
Probabilistic Inference and Differential Privacy
Oliver Williams, Frank Mcsherry
We identify and investigate a strong connection between probabilistic inference and differential privacy, the latter being a recent privacy definition that permits only indirect observation of data through noisy measurement. Previous research on differential privacy has focused on designing measurement processes whose output is likely to be useful on its own. We consider the potential of applying probabilistic inference to the measurements and measurement process to derive posterior distributions over the data sets and model parameters thereof. We find that probabilistic inference can improve accuracy, integrate multiple observations, measure uncertainty, and even provide posterior distributions over quantities that were not directly measured.
****
Copula Processes
Andrew G. Wilson, Zoubin Ghahramani
We define a copula process which describes the dependencies between arbitrarily many random variables independently of their marginal distributions. As an example, we develop a stochastic volatility model, Gaussian Copula Process Volatility (GCPV), to predict the latent standard deviations of a sequence of random variables. To make predictions we use Bayesian inference, with the Laplace approximation, and with Markov chain Monte Carlo as an alternative. We find our model can outperform GARCH on simulated and financial data. And unlike GARCH, GCPV can easily handle missing data, incorporate covariates other than time, and model a rich class of covariance structures.
****

Learning invariant features using the Transformed Indian Buffet Process
Joseph Austerweil, Thomas Griffiths
Identifying the features of objects becomes a challenge when those features can change in their appearance. We introduce the Transformed Indian Buffet Process (tIBP), and use it to define a nonparametric Bayesian model that infers features that can transform across instantiations. We show that this model can identify features that are location invariant by modeling a previous experiment on human feature learning. However, allowing features to transform adds new kinds of ambiguity: Are two parts of an object the same feature with different transformations or two unique features? What transformations can features undergo? We present two new experiments in which we explore how people resolve these questions, showing that the tIBP model demonstrates a similar sensitivity to context to that shown by human learners when determining the invariant aspects of features.
****
An Inverse Power Method for Nonlinear Eigenproblems with Applications in 1-Spectral Clustering and Sparse PCA
Matthias Hein, Thomas Bühler
Many problems in machine learning and statistics can be formulated as (generalized) eigenproblems. In terms of the associated optimization problem, computing linear eigenvectors amounts to finding critical points of a quadratic function subject to quadratic constraints. In this paper we show that a certain class of constrained optimization problems with nonquadratic objective and constraints can be understood as nonlinear eigenproblems. We derive a generalization of the inverse power method which is guaranteed to converge to a nonlinear eigenvector. We apply the inverse power method to 1-spectral clustering and sparse PCA which can naturally be formulated as nonlinear eigenproblems. In both applications we achieve state-of-the-art results in terms of solution quality and runtime. Moving beyond the standard eigenproblem should be useful also in many other applications and our inverse power method can be easily adapted to new problems.
****
Fast detection of multiple change-points shared by many signals using group LARS
Jean-philippe Vert, Kevin Bleakley
We present a fast algorithm for the detection of multiple change-points when each is frequently shared by members of a set of co-occurring one-dimensional signals. We give conditions on consistency of the method when the number of signals increases, and provide empirical evidence to support the consistency results.
****
Learning To Count Objects in Images
Victor Lempitsky, Andrew Zisserman
We propose a new supervised learning framework for visual object counting tasks, such as estimating the number of cells in a microscopic image or the number of humans in surveillance video frames. We focus on the practically-attractive case when the training images are annotated with dots (one dot per object). Our goal is to accurately estimate the count. However, we evade the hard task of learning to detect and localize individual object instances. Instead, we cast the problem as that of estimating an image density whose integral over any image region gives the count of objects within that region. Learning to infer such density can be formulated as a minimization of a regularized risk quadratic cost function. We introduce a new loss function, which is well-suited for such learning, and at the same time can be computed efficiently via a maximum subarray algorithm. The learning can then be posed as a convex quadratic program solvable with cutting-plane optimization. The proposed framework is very flexible as it can accept any domain-specific visual features. Once trained, our system provides accurate object counts and requires a very small time overhead over the feature extraction step, making it a good candidate for applications involving real-time processing or dealing with huge amount of visual data.
****
Robust PCA via Outlier Pursuit
Huan Xu, Constantine Caramanis, Sujay Sanghavi
Singular Value Decomposition (and Principal Component Analysis) is one of the m

ost widely used techniques for dimensionality reduction: successful and efficien tly computable, it is nevertheless plagued by a well-known, well-documented sens itivity to outliers. Recent work has considered the setting where each point has a few arbitrarily corrupted components. Yet, in applications of SVD or PCA such as robust collaborative filtering or bioinformatics, malicious agents, defectiv e genes, or simply corrupted or contaminated experiments may effectively yield e ntire points that are completely corrupted. We present an efficient convex opti mization-based algorithm we call Outlier Pursuit, that under some mild assumpti ons on the uncorrupted points (satisfied, e.g., by the standard generative assum ption in PCA problems) recovers the exact optimal low-dimensional subspace, and identifies the corrupted points. Such identification of corrupted points that do not conform to the low-dimensional approximation, is of paramount interest in b ioinformatics and financial applications, and beyond. Our techniques involve mat rix decomposition using nuclear norm minimization, however, our results, setup, and approach, necessarily differ considerably from the existing line of work in matrix completion and matrix decomposition, since we develop an approach to reco ver the correct column space of the uncorrupted matrix, rather than the exact ma trix itself.
****

Multi-label Multiple Kernel Learning by Stochastic Approximation: Application to Visual Object Recognition

Serhat Bucak, Rong Jin, Anil Jain

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

Learning sparse dynamic linear systems using stable spline kernels and exponenti al hyperpriors

Alessandro Chiuso, Gianluigi Pillonetto

We introduce a new Bayesian nonparametric approach to identification of sparse d ynamic linear systems. The impulse responses are modeled as Gaussian processes w hose autocovariances encode the BIBO stability constraint, as defined by the rec ently introduced “Stable Spline kernel”. Sparse solutions are obtained by placin g exponential hyperpriors on the scale factors of such kernels. Numerical experi ments regarding estimation of ARMAX models show that this technique provides a d efinite advantage over a group LAR algorithm and state-of-the-art parametric ide ntification techniques based on prediction error minimization.
****