Near-minimax recursive density estimation on the binary hypercube
Maxim Raginsky, Svetlana Lazebnik, Rebecca Willett, Jorge Silva
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.
****

Translated Learning: Transfer Learning across Different Feature Spaces
Wenyuan Dai, Yuqiang Chen, Gui-rong Xue, Qiang Yang, Yong Yu
This paper investigates a new machine learning strategy called translated learning. Unlike many previous learning tasks, we focus on how to use labeled data from one feature space to enhance the classification of other entirely different learning spaces. For example, we might wish to use labeled text data to help learn a model for classifying image data, when the labeled images are difficult to obtain. An important aspect of translated learning is to build a "bridge" to link one feature space (known as the "source space") to another space (known as the "target space") through a translator in order to migrate the knowledge from source to target. The translated learning solution uses a language model to link the class labels to the features in the source spaces, which in turn is translated to the features in the target spaces. Finally, this chain of linkages is completed by tracing back to the instances in the target spaces. We show that this path of linkage can be modeled using a Markov chain and risk minimization. Through experiments on the text-aided image classification and cross-language classification tasks, we demonstrate that our translated learning framework can greatly outperform many state-of-the-art baseline methods.
****

Learning Bounded Treewidth Bayesian Networks
Gal Elidan, Stephen Gould
With the increased availability of data for complex domains, it is desirable to learn Bayesian network structures that are sufficiently expressive for generalization while also allowing for tractable inference. While the method of thin junction trees can, in principle, be used for this purpose, its fully greedy nature makes it prone to overfitting, particularly when data is scarce. In this work we present a novel method for learning Bayesian networks of bounded treewidth that employs global structure modifications and that is polynomial in the size of the graph and the treewidth bound. At the heart of our method is a triangulated graph that we dynamically update in a way that facilitates the addition of chain structures that increase the bound on the model's treewidth by at most one. We demonstrate the effectiveness of our ``treewidth-friendly'' method on several real-life datasets. Importantly, we also show that by using global operators, we are able to achieve better generalization even when learning Bayesian networks of unbounded treewidth.
****

Local Gaussian Process Regression for Real Time Online Model Learning
Duy Nguyen-tuong, Jan Peters, Matthias Seeger
Learning in real-time applications, e.g., online approximation of the inverse dynamics model for model-based robot control, requires fast online regression techniques. Inspired by local learning, we propose a method to speed up standard Gaussian Process regression (GPR) with local GP models (LGP). The training data is partitioned in local regions, for each an individual GP model is trained. The prediction for a query point is performed by weighted estimation using nearby local models. Unlike other GP approximations, such as mixtures of experts, we use a distance based measure for partitioning of the data and weighted prediction. The proposed method achieves online learning and prediction in real-time. Comparisons with other nonparametric regression methods show that LGP has higher accuracy than LWPR and close to the performance of standard GPR and nu-SVR.
****

Grouping Contours Via a Related Image
Praveen Srinivasan, Liming Wang, Jianbo Shi
Contours have been established in the biological and computer vision literatures

as a compact yet descriptive representation of object shape. While individual c
ontours provide structure, they lack the large spatial support of region segment
s (which lack internal structure). We present a method for further grouping of c
ontours in an image using their relationship to the contours of a second, relate
d image. Stereo, motion, and similarity all provide cues that can aid this task;
 contours that have similar transformations relating them to their matching cont
ours in the second image likely belong to a single group. To find matches for co
ntours, we rely only on shape, which applies directly to all three modalities wi
thout modification, in constrant to the specialized approaches developed for eac
h independently. Visually salient contours are extracted in each image, along wi
th a set of candidate transformations for aligning subsets of them. For each tra
nsformation, groups of contours with matching shape across the two images are id
entified to provide a context for evaluating matches of individual contour point
s across the images. The resulting contexts of contours are used to perform a fi
nal grouping on contours in the original image while simultaneously finding matc
hes in the related image, again by shape matching. We demonstrate grouping resul
ts on image pairs consisting of stereo, motion, and similar images. Our method a
lso produces qualitatively better results against a baseline method that does no
t use the inferred contexts.
****

Designing neurophysiology experiments to optimally constrain receptive field mod
els along parametric submanifolds
Jeremy Lewi, Robert Butera, David Schneider, Sarah Woolley, Liam Paninski
Sequential optimal design methods hold great promise for improving the efficienc
y of neurophysiology experiments. However, previous methods for optimal experime
ntal design have incorporated only weak prior information about the underlying n
eural system (e.g., the sparseness or smoothness of the receptive field). Here w
e describe how to use stronger prior information, in the form of parametric mode
ls of the receptive field, in order to construct optimal stimuli and further imp
rove the efficiency of our experiments. For example, if we believe that the rece
ptive field is well-approximated by a Gabor function, then our method constructs
 stimuli that optimally constrain the Gabor parameters (orientation, spatial fre
quency, etc.) using as few experimental trials as possible. More generally, we m
ay believe a priori that the receptive field lies near a known sub-manifold of t
he full parameter space; in this case, our method chooses stimuli in order to re
duce the uncertainty along the tangent space of this sub-manifold as rapidly as
possible. Applications to simulated and real data indicate that these methods ma
y in many cases improve the experimental efficiency.
****

Analyzing human feature learning as nonparametric Bayesian inference
Thomas Griffiths, Joseph Austerweil
Almost all successful machine learning algorithms and cognitive models require p
owerful representations capturing the features that are relevant to a particular
 problem. We draw on recent work in nonparametric Bayesian statistics to define
a rational model of human feature learning that forms a featural representation
from raw sensory data without pre-specifying the number of features. By comparin
g how the human perceptual system and our rational model use distributional and
category information to infer feature representations, we seek to identify some
of the forces that govern the process by which people separate and combine senso
ry primitives to form features.
****

From Online to Batch Learning with Cutoff-Averaging
Ofer Dekel
We present cutoff averaging", a technique for converting any conservative online
 learning algorithm into a batch learning algorithm. Most online-to-batch conver
sion techniques work well with certain types of online learning algorithms and n
ot with others, whereas cutoff averaging explicitly tries to adapt to the charac
teristics of the online algorithm being converted. An attractive property of our
 technique is that it preserves the efficiency of the original online algorithm,
 making it approporiate for large-scale learning problems. We provide a statisti

cal analysis of our technique and back our theoretical claims with experimental results."
****
Structured ranking learning using cumulative distribution networks
Jim Huang, Brendan J. Frey
Ranking is at the heart of many information retrieval applications. Unlike standard regression or classification, in which we predict outputs independently, in ranking, we are interested in predicting structured outputs so that misranking one object can significantly affect whether we correctly rank the other objects. In practice, the problem of ranking involves a large number of objects to be ranked and either approximate structured prediction methods are required, or assumptions of independence between object scores must be made in order to make the problem tractable. We present a probabilistic method for learning to rank using the graphical modelling framework of cumulative distribution networks (CDNs), where we can take into account the structure inherent to the problem of ranking by modelling the joint cumulative distribution functions (CDFs) over multiple pairwise preferences. We apply our framework to the problem of document retrieval in the case of the OHSUMED benchmark dataset. We will show that the RankNet, ListNet and ListMLE probabilistic models can be viewed as particular instances of CDNs and that our proposed framework allows for the exploration of a broad class of flexible structured loss functionals for ranking learning.
****
Asynchronous Distributed Learning of Topic Models
Padhraic Smyth, Max Welling, Arthur Asuncion
Distributed learning is a problem of fundamental interest in machine learning and cognitive science. In this paper, we present asynchronous distributed learning algorithms for two well-known unsupervised learning frameworks: Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Processes (HDP). In the proposed approach, the data are distributed across P processors, and processors independently perform Gibbs sampling on their local data and communicate their information in a local asynchronous manner with other processors. We demonstrate that our asynchronous algorithms are able to learn global topic models that are statistically as accurate as those learned by the standard LDA and HDP samplers, but with significant improvements in computation time and memory. We show speedup results on a 730-million-word text corpus using 32 processors, and we provide perplexity results for up to 1500 virtual processors. As a stepping stone in the development of asynchronous HDP, a parallel HDP sampler is also introduced.
****
Cascaded Classification Models: Combining Models for Holistic Scene Understanding
Geremy Heitz, Stephen Gould, Ashutosh Saxena, Daphne Koller
One of the original goals of computer vision was to fully understand a natural scene. This requires solving several problems simultaneously, including object detection, labeling of meaningful regions, and 3d reconstruction. While great progress has been made in tackling each of these problems in isolation, only recently have researchers again been considering the difficult task of assembling various methods to the mutual benefit of all. We consider learning a set of such classification models in such a way that they both solve their own problem and help each other. We develop a framework known as Cascaded Classification Models (CCM), where repeated instantiations of these classifiers are coupled by their input/output variables in a cascade that improves performance at each level. Our method requires only a limited â■■black boxâ■■ interface with the models, allowing us to use very sophisticated, state-of-the-art classifiers without having to look under the hood. We demonstrate the effectiveness of our method on a large set of natural images by combining the subtasks of scene categorization, object detection, multiclass image segmentation, and 3d scene reconstruction.
****
Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes
Ben Calderhead, Mark Girolami, Neil Lawrence

Identification and comparison of nonlinear dynamical systems using noisy and spa
rse experimental data is a vital task in many fields, however current methods ar
e computationally expensive and prone to error due in part to the nonlinear natu
re of the likelihood surfaces induced. We present an accelerated sampling proced
ure which enables Bayesian inference of parameters in nonlinear ordinary and del
ay differential equations via the novel use of Gaussian processes (GP). Our meth
od involves GP regression over time-series data, and the resulting derivative an
d time delay estimates make parameter inference possible without solving the dyn
amical system explicitly, resulting in dramatic savings of computational time. W
e demonstrate the speed and statistical accuracy of our approach using examples
of both ordinary and delay differential equations, and provide a comprehensive c
omparison with current state of the art methods.
****

Linear Classification and Selective Sampling Under Low Noise Conditions
Giovanni Cavallanti, Nicolò Cesa-bianchi, Claudio Gentile
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.
****

On the Efficient Minimization of Classification Calibrated Surrogates
Richard Nock, Frank Nielsen
Bartlett et al (2006) recently proved that a ground condition for convex surroga
tes, classification calibration, ties up the minimization of the surrogates and
classification risks, and left as an important problem the algorithmic questions
 about the minimization of these surrogates. In this paper, we propose an algori
thm which provably minimizes any classification calibrated surrogate strictly co
nvex and differentiable --- a set whose losses span the exponential, logistic an
d squared losses ---, with boosting-type guaranteed convergence rates under a we
ak learning assumption. A particular subclass of these surrogates, that we call
balanced convex surrogates, has a key rationale that ties it to maximum likeliho
od estimation, zero-sum games and the set of losses that satisfy some of the mos
t common requirements for losses in supervised learning. We report experiments o
n more than 50 readily available domains of 11 flavors of the algorithm, that sh
ed light on new surrogates, and the potential of data dependent strategies to tu
ne surrogates.
****

Unlabeled data: Now it helps, now it doesn't
Aarti Singh, Robert Nowak, Jerry Zhu
Empirical evidence shows that in favorable situations semi-supervised learning (
SSL) algorithms can capitalize on the abundancy of unlabeled training data to im
prove the performance of a learning task, in the sense that fewer labeled traini
ng data are needed to achieve a target error bound. However, in other situations
 unlabeled data do not seem to help. Recent attempts at theoretically characteri
zing the situations in which unlabeled data can help have met with little succes
s, and sometimes appear to conflict with each other and intuition. In this paper
, we attempt to bridge the gap between practice and theory of semi-supervised le
arning. We develop a rigorous framework for analyzing the situations in which un
labeled data can help and quantify the improvement possible using finite sample
error bounds. We show that there are large classes of problems for which SSL can
 significantly outperform supervised learning, in finite sample regimes and some
times also in terms of error convergence rates.
****

Unifying the Sensory and Motor Components of Sensorimotor Adaptation
Adrian Haith, Carl Jackson, R. Miall, Sethu Vijayakumar
Adaptation of visually guided reaching movements in novel visuomotor environment
s (e.g. wearing prism goggles) comprises not only motor adaptation but also subs
tantial sensory adaptation, corresponding to shifts in the perceived spatial loc
ation of visual and proprioceptive cues. Previous computational models of the se
nsory component of visuomotor adaptation have assumed that it is driven purely b

y the discrepancy introduced between visual and proprioceptive estimates of hand position and is independent of any motor component of adaptation. We instead propose a unified model in which sensory and motor adaptation are jointly driven by optimal Bayesian estimation of the sensory and motor contributions to perceived errors. Our model is able to account for patterns of performance errors during visuomotor adaptation as well as the subsequent perceptual aftereffects. This unified model also makes the surprising prediction that force field adaptation will elicit similar perceptual shifts, even though there is never any discrepancy between visual and proprioceptive observations. We confirm this prediction with an experiment.

****

Modeling human function learning with Gaussian processes

Thomas Griffiths, Chris Lucas, Joseph Williams, Michael Kalish

Accounts of how people learn functional relationships between continuous variables have tended to focus on two possibilities: that people are estimating explicit functions, or that they are simply performing associative learning supported by similarity. We provide a rational analysis of function learning, drawing on work on regression in machine learning and statistics. Using the equivalence of Bayesian linear regression and Gaussian processes, we show that learning explicit rules and using similarity can be seen as two views of one solution to this problem. We use this insight to define a Gaussian process model of human function learning that combines the strengths of both approaches.

****

Hebbian Learning of Bayes Optimal Decisions

Bernhard Nessler, Michael Pfeiffer, Wolfgang Maass

Uncertainty is omnipresent when we perceive or interact with our environment, and the Bayesian framework provides computational methods for dealing with it. Mathematical models for Bayesian decision making typically require datastructures that are hard to implement in neural networks. This article shows that even the simplest and experimentally best supported type of synaptic plasticity, Hebbian learning, in combination with a sparse, redundant neural code, can in principle learn to infer optimal Bayesian decisions. We present a concrete Hebbian learning rule operating on log-probability ratios. Modulated by reward-signals, this Hebbian plasticity rule also provides a new perspective for understanding how Bayesian inference could support fast reinforcement learning in the brain. In particular we show that recent experimental results by Yang and Shadlen [1] on reinforcement learning of probabilistic inference in primates can be modeled in this way.

****

An Empirical Analysis of Domain Adaptation Algorithms for Genomic Sequence Analysis

Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, Bernhard Schölkopf

We study the problem of domain transfer for a supervised classification task in mRNA splicing. We consider a number of recent domain transfer methods from machine learning, including some that are novel, and evaluate them on genomic sequence data from model organisms of varying evolutionary distance. We find that in cases where the organisms are not closely related, the use of domain adaptation methods can help improve classification performance.

****

Kernel Change-point Analysis

Zaïd Harchaoui, Eric Moulines, Francis Bach

We introduce a kernel-based method for change-point analysis within a sequence of temporal observations. Change-point analysis of an (unlabelled) sample of observations consists in, first, testing whether a change in the distribution occurs within the sample, and second, if a change occurs, estimating the change-point instant after which the distribution of the observations switches from one distribution to another different distribution. We propose a test statistics based upon the maximum kernel Fisher discriminant ratio as a measure of homogeneity between segments. We derive its limiting distribution under the null hypothesis (no change occurs), and establish the consistency under the alternative hypothesis (

a change occurs). This allows to build a statistical hypothesis testing procedur
e for testing the presence of change-point, with a prescribed false-alarm probab
ility and detection probability tending to one in the large-sample setting. If a
 change actually occurs, the test statistics also yields an estimator of the cha
nge-point location. Promising experimental results in temporal segmentation of m
ental tasks from BCI data and pop song indexation are presented.
****

Biasing Approximate Dynamic Programming with a Lower Discount Factor
Marek Petrik, Bruno Scherrer
Most algorithms for solving Markov decision processes rely on a discount factor,
 which ensures their convergence. In fact, it is often used in problems with is
no intrinsic motivation. In this paper, we show that when used in approximate dy
namic programming, an artificially low discount factor may significantly improve
 the performance on some problems, such as Tetris. We propose two explanations f
or this phenomenon. Our first justification follows directly from the standard a
pproximation error bounds: using a lower discount factor may decrease the approx
imation error bounds. However, we also show that these bounds are loose, a thus
their decrease does not entirely justify a better practical performance. We thus
 propose another justification: when the rewards are received only sporadically
(as it is the case in Tetris), we can derive tighter bounds, which support a sig
nificant performance increase with a decrease in the discount factor.
****

Performance analysis for L_2 kernel classification
Jooseuk Kim, Clayton Scott
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.
****

Influence of graph construction on graph-based clustering measures
Markus Maier, Ulrike Luxburg, Matthias Hein
Graph clustering methods such as spectral clustering are defined for general wei
ghted graphs. In machine learning, however, data often is not given in form of a
 graph, but in terms of similarity (or distance) values between points. In this
case, first a neighborhood graph is constructed using the similarities between t
he points and then a graph clustering algorithm is applied to this graph. In thi
s paper we investigate the influence of the construction of the similarity graph
 on the clustering results. We first study the convergence of graph clustering c
riteria such as the normalized cut (Ncut) as the sample size tends to infinity.
We find that the limit expressions are different for different types of graph, f
or example the r-neighborhood graph or the k -nearest neighbor graph. In plain w
ords: Ncut on a kNN graph does something systematically different than Ncut on a
n r-neighborhood graph! This finding shows that graph clustering criteria cannot
 be studied independently of the kind of graph they are applied to. We also prov
ide examples which show that these differences can be observed for toy and real
data already for rather small sample sizes.
****

MDPs with Non-Deterministic Policies
M. Fard, Joelle Pineau
Markov Decision Processes (MDPs) have been extensively studied and used in the c
ontext of planning and decision-making, and many methods exist to find the optim
al policy for problems modelled as MDPs. Although finding the optimal policy is
sufficient in many domains, in certain applications such as decision support sys
tems where the policy is executed by a human (rather than a machine), finding al
l possible near-optimal policies might be useful as it provides more flexibility
 to the person executing the policy. In this paper we introduce the new concept
of non-deterministic MDP policies, and address the question of finding near-opti
mal non-deterministic policies. We propose two solutions to this problem, one ba
sed on a Mixed Integer Program and the other one based on a search algorithm. We
 include experimental results obtained from applying this framework to optimize

treatment choices in the context of a medical decision support system.
****
Semi-supervised Learning with Weakly-Related Unlabeled Data : Towards Better Tex t Categorization
Liu Yang, Rong Jin, Rahul Sukthankar
The cluster assumption is exploited by most semi-supervised learning (SSL) metho ds. However, if the unlabeled data is merely weakly related to the target classe s, it becomes questionable whether driving the decision boundary to the low dens ity regions of the unlabeled data will help the classification. In such case, th e cluster assumption may not be valid; and consequently how to leverage this typ e of unlabeled data to enhance the classification accuracy becomes a challenge. We introduce Semi-supervised Learning with Weakly-Related Unlabeled Data" (SSLW) , an inductive method that builds upon the maximum-margin approach, towards a be tter usage of weakly-related unlabeled information. Although the SSLW could impr ove a wide range of classification tasks, in this paper, we focus on text catego rization with a small training pool. The key assumption behind this work is that , even with different topics, the word usage patterns across different corpora t ends to be consistent. To this end, SSLW estimates the optimal word-correlation matrix that is consistent with both the co-occurrence information derived from t he weakly-related unlabeled documents and the labeled documents. For empirical e valuation, we present a direct comparison with a number of state-of-the-art meth ods for inductive semi-supervised learning and text categorization; and we show that SSLW results in a significant improvement in categorization accuracy, equip ped with a small training set and an unlabeled resource that is weakly related t o the test beds."
****
Evaluating probabilities under high-dimensional latent variable models
Iain Murray, Russ R. Salakhutdinov
We present a simple new Monte Carlo algorithm for evaluating probabilities of ob servations in complex latent variable models, such as Deep Belief Networks. Whil e the method is based on Markov chains, estimates based on short runs are formal ly unbiased. In expectation, the log probability of a test set will be underesti mated, and this could form the basis of a probabilistic bound. The method is muc h cheaper than gold-standard annealing-based methods and only slightly more expe nsive than the cheapest Monte Carlo methods. We give examples of the new method substantially improving simple variational bounds at modest extra cost.
****
Counting Solution Clusters in Graph Coloring Problems Using Belief Propagation
Lukas Kroc, Ashish Sabharwal, Bart Selman
We show that an important and computationally challenging solution space feature of the graph coloring problem (COL), namely the number of clusters of solutions , can be accurately estimated by a technique very similar to one for counting th e number of solutions. This cluster counting approach can be naturally written i n terms of a new factor graph derived from the factor graph representing the COL instance. Using a variant of the Belief Propagation inference framework, we can ef■ciently approximate cluster counts in random COL problems over a large range of graph densities. We illustrate the algorithm on instances with up to 100, 00 0 vertices. Moreover, we supply a methodology for computing the number of clus- ters exactly using advanced techniques from the knowledge compilation literature . This methodology scales up to several hundred variables.
****
Supervised Bipartite Graph Inference
Yoshihiro Yamanishi
We formulate the problem of bipartite graph inference as a supervised learning p roblem, and propose a new method to solve it from the viewpoint of distance metr ic learning. The method involves the learning of two mappings of the heterogeneo us objects to a unified Euclidean space representing the network topology of the bipartite graph, where the graph is easy to infer. The algorithm can be formula ted as an optimization problem in a reproducing kernel Hilbert space. We report encouraging results on the problem of compound-protein interaction network recon

struction from chemical structure data and genomic sequence data.
****
Domain Adaptation with Multiple Sources
Yishay Mansour, Mehryar Mohri, Afshin Rostamizadeh
This paper presents a theoretical analysis of the problem of adaptation with multiple sources. For each source domain, the distribution over the input points as well as a hypothesis with error at most \epsilon are given. The problem consists of combining these hypotheses to derive a hypothesis with small error with respect to the target domain. We present several theoretical results relating to this problem. In particular, we prove that standard convex combinations of the source hypotheses may in fact perform very poorly and that, instead, combinations weighted by the source distributions benefit from favorable theoretical guarantees. Our main result shows that, remarkably, for any fixed target function, there exists a distribution weighted combining rule that has a loss of at most \epsilon with respect to any target mixture of the source distributions. We further generalize the setting from a single target function to multiple consistent target functions and show the existence of a combining rule with error at most 3\epsilon. Finally, we report empirical results for a multiple source adaptation problem with a real-world dataset.
****
Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning
Ali Rahimi, Benjamin Recht
Randomized neural networks are immortalized in this AI Koan: In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6. What are you doing?'' asked Minsky.I am training a randomly wired neural net to play tic-tac-toe,'' Sussman replied. Why is the net wired randomly?'' asked Minsky. Sussman replied,I do not want it to have any preconceptions of how to play.'' Minsky then shut his eyes. Why do you close your eyes?'' Sussman asked his teacher.So that the room will be empty,'' replied Minsky. At that moment, Sussman was enlightened. We analyze shallow random networks with the help of concentration of measure inequalities. Specifically, we consider architectures that compute a weighted sum of their inputs after passing them through a bank of arbitrary randomized nonlinearities. We identify conditions under which these networks exhibit good classification performance, and bound their test error in terms of the size of the dataset and the number of random nonlinearities.
****
Regularized Learning with Networks of Features
Ted Sandler, John Blitzer, Partha Talukdar, Lyle Ungar
For many supervised learning problems, we possess prior knowledge about which features yield similar information about the target variable. In predicting the topic of a document, we might know that two words are synonyms, or when performing image recognition, we know which pixels are adjacent. Such synonymous or neighboring features are near-duplicates and should therefore be expected to have similar weights in a good model. Here we present a framework for regularized learning in settings where one has prior knowledge about which features are expected to have similar and dissimilar weights. This prior knowledge is encoded as a graph whose vertices represent features and whose edges represent similarities and dissimilarities between them. During learning, each feature's weight is penalized by the amount it differs from the average weight of its neighbors. For text classification, regularization using graphs of word co-occurrences outperforms manifold learning and compares favorably to other recently proposed semi-supervised learning methods. For sentiment analysis, feature graphs constructed from declarative human knowledge, as well as from auxiliary task learning, significantly improve prediction accuracy.
****
Breaking Audio CAPTCHAs
Jennifer Tam, Jiri Simsa, Sean Hyde, Luis Ahn
CAPTCHAs are computer-generated tests that humans can pass but current computer systems cannot. CAPTCHAs provide a method for automatically distinguis

hing a human from a computer program, and therefore can protect Web services fro m abuse by so-called "bots." Most CAP T C H A s consist of distorted images, usu ally text, for which a user must provide some description. Unfortunately, visual CAP T C H A s limit access to the millions of visually impaired people using th e Web. Audio CAP T C H A s were created to solve this accessibility issue; howev er, the security of audio CAP T C H A s was never formally tested. Some visual C AP T C H A s have been broken using machine learning techniques, and we propose using similar ideas to test the security of audio CAP T C H A s . Audio CAP T C H A s are generally composed of a set of words to be identified, layered on top of noise. We analyzed the security of current audio CAP T CH A s from popular We b sites by using AdaBoost, SVM, and k-NN, and achieved correct solutions for tes t samples with accuracy up to 71%. Such accuracy is enough to consider these CAP TCHAs broken. Training several different machine learning algorithms on differen t types of audio CAP T C H A s allowed us to analyze the strengths and weaknesse s of the algorithms so that we could suggest a design for a more robust audio CA PTCHA.

****

Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Ou tlier Detection

Takafumi Kanamori, Shohei Hido, Masashi Sugiyama

We address the problem of estimating the ratio of two probability density functi ons (a.k.a.~the importance). The importance values can be used for various succe eding tasks such as non-stationarity adaptation or outlier detection. In this pa per, we propose a new importance estimation method that has a closed-form soluti on; the leave-one-out cross-validation score can also be computed analytically. Therefore, the proposed method is computationally very efficient and numerically stable. We also elucidate theoretical properties of the proposed method such as the convergence rate and approximation error bound. Numerical experiments show that the proposed method is comparable to the best existing method in accuracy, while it is computationally more efficient than competing approaches.

****

Differentiable Sparse Coding

J. Bagnell, David Bradley

Prior work has shown that features which appear to be biologically plausible as well as empirically useful can be found by sparse coding with a prior such as a laplacian (L1) that promotes sparsity. We show how smoother priors can pre- serv e the bene■ts of these sparse priors while adding stability to the Maximum A-Pos teriori (MAP) estimate that makes it more useful for prediction problems. Additi onally, we show how to calculate the derivative of the MAP estimate ef■- ciently with implicit differentiation. One prior that can be differentiated this way is KL-regularization. We demonstrate its effectiveness on a wide variety of appli- cations, and ■nd that online optimization of the parameters of the KL-regulariz ed model can signi■cantly improve prediction performance.

****

Inferring rankings under constrained sensing

Srikanth Jagabathula, Devavrat Shah

Motivated by applications like elections, web-page ranking, revenue maximization etc., we consider the question of inferring popular rankings using constrained data. More specifically, we consider the problem of inferring a probability dist ribution over the group of permutations using its first order marginals. We firs t prove that it is not possible to recover more than O(n) permutations over n el ements with the given information. We then provide a simple and novel algorithm that can recover up to O(n) permutations under a natural stochastic model; in th is sense, the algorithm is optimal. In certain applications, the interest is in recovering only the most popular (or mode) ranking. As a second result, we provi de an algorithm based on the Fourier Transform over the symmetric group to recov er the mode under a natural majority condition; the algorithm turns out to be a maximum weight matching on an appropriately defined weighted bipartite graph. Th e questions considered are also thematically related to Fourier Transforms over the symmetric group and the currently popular topic of compressed sensing.

****
Sparse Convolved Gaussian Processes for Multi-output Regression
Mauricio Alvarez, Neil Lawrence
We present a sparse approximation approach for dependent output Gaussian processes (GP). Employing a latent function framework, we apply the convolution process formalism to establish dependencies between output variables, where each latent function is represented as a GP. Based on these latent functions, we establish an approximation scheme using a conditional independence assumption between the output processes, leading to an approximation of the full covariance which is determined by the locations at which the latent functions are evaluated. We show results of the proposed methodology for synthetic data and real world applications on pollution prediction and a sensor network.
****
A ``Shape Aware'' Model for semi-supervised Learning of Objects and its Context
Abhinav Gupta, Jianbo Shi, Larry S. Davis
Integrating semantic and syntactic analysis is essential for document analysis. Using an analogous reasoning, we present an approach that combines bag-of-words and spatial models to perform semantic and syntactic analysis for recognition of an object based on its internal appearance and its context. We argue that while object recognition requires modeling relative spatial locations of image features within the object, a bag-of-word is sufficient for representing context. Learning such a model from weakly labeled data involves labeling of features into two classes: foreground(object) or ''informative'' background(context). labeling. We present a ''shape-aware'' model which utilizes contour information for efficient and accurate labeling of features in the image. Our approach iterates between an MCMC-based labeling and contour based labeling of features to integrate co-occurrence of features and shape similarity.
****
Multi-task Gaussian Process Learning of Robot Inverse Dynamics
Christopher Williams, Stefan Klanke, Sethu Vijayakumar, Kian Chai
The inverse dynamics problem for a robotic manipulator is to compute the torques needed at the joints to drive it along a given trajectory; it is beneficial to be able to learn this function for adaptive control. A given robot manipulator will often need to be controlled while holding different loads in its end effector, giving rise to a multi-task learning problem. We show how the structure of the inverse dynamics problem gives rise to a multi-task Gaussian process prior over functions, where the inter-task similarity depends on the underlying dynamic parameters. Experiments demonstrate that this multi-task formulation generally improves performance over either learning only on single tasks or pooling the data over all tasks.
****
Efficient Inference in Phylogenetic InDel Trees
Alexandre Bouchard-côté, Dan Klein, Michael Jordan
Accurate and efficient inference in evolutionary trees is a central problem in computational biology. Realistic models require tracking insertions and deletions along the phylogenetic tree, making inference challenging. We propose new sampling techniques that speed up inference and improve the quality of the samples. We compare our method to previous approaches and show performance improvement on metrics evaluating multiple sequence alignment and reconstruction of ancestral sequences.
****
Estimating vector fields using sparse basis field expansions
Stefan Haufe, Vadim Nikulin, Andreas Ziehe, Klaus-Robert Müller, Guido Nolte
We introduce a novel framework for estimating vector fields using sparse basis field expansions (S-FLEX). The notion of basis fields, which are an extension of scalar basis functions, arises naturally in our framework from a rotational invariance requirement. We consider a regression setting as well as inverse problems. All variants discussed lead to second-order cone programming formulations. While our framework is generally applicable to any type of vector field, we focus in this paper on applying it to solving the EEG/MEG inverse problem. It is shown

that significantly more precise and neurophysiologically more plausible location and shape estimates of cerebral current sources from EEG/MEG measurements become possible with our method when comparing to the state-of-the-art.

****

Probabilistic detection of short events, with application to critical care monitoring

Norm Aleks, Stuart J. Russell, Michael Madden, Diane Morabito, Kristan Staudenmayer, Mitchell Cohen, Geoffrey Manley

We describe an application of probabilistic modeling and inference technology to the problem of analyzing sensor data in the setting of an intensive care unit (ICU). In particular, we consider the arterial-line blood pressure sensor, which is subject to frequent data artifacts that cause false alarms in the ICU and make the raw data almost useless for automated decision making. The problem is complicated by the fact that the sensor data are acquired at fixed intervals whereas the events causing data artifacts may occur at any time and have durations that may be significantly shorter than the data collection interval. We show that careful modeling of the sensor, combined with a general technique for detecting sub-interval events and estimating their duration, enables effective detection of artifacts and accurate estimation of the underlying blood pressure values.

****

Spectral Clustering with Perturbed Data

Ling Huang, Donghui Yan, Nina Taft, Michael Jordan

Spectral clustering is useful for a wide-ranging set of applications in areas such as biological data analysis, image processing and data mining. However, the computational and/or communication resources required by the method in processing large-scale data sets are often prohibitively high, and practitioners are often required to perturb the original data in various ways (quantization, downsampling, etc) before invoking a spectral algorithm. In this paper, we use stochastic perturbation theory to study the effects of data perturbation on the performance of spectral clustering. We show that the error under perturbation of spectral clustering is closely related to the perturbation of the eigenvectors of the Laplacian matrix. From this result we derive approximate upper bounds on the clustering error. We show that this bound is tight empirically across a wide range of problems, suggesting that it can be used in practical settings to determine the amount of data reduction allowed in order to meet a specification of permitted loss in clustering performance.

****

Estimating the Location and Orientation of Complex, Correlated Neural Activity using MEG

Julia Owen, Hagai Attias, Kensuke Sekihara, Srikantan Nagarajan, David Wipf

The synchronous brain activity measured via MEG (or EEG) can be interpreted as arising from a collection (possibly large) of current dipoles or sources located throughout the cortex. Estimating the number, location, and orientation of these sources remains a challenging task, one that is significantly compounded by the effects of source correlations and the presence of interference from spontaneous brain activity, sensor noise, and other artifacts. This paper derives an empirical Bayesian method for addressing each of these issues in a principled fashion. The resulting algorithm guarantees descent of a cost function uniquely designed to handle unknown orientations and arbitrary correlations. Robust interference suppression is also easily incorporated. In a restricted setting, the proposed method is shown to have theoretically zero bias estimating both the location and orientation of multi-component dipoles even in the presence of correlations, unlike a variety of existing Bayesian localization methods or common signal processing techniques such as beamforming and sLORETA. Empirical results on both simulated and real data sets verify the efficacy of this approach.

****

Multi-stage Convex Relaxation for Learning with Sparse Regularization

Tong Zhang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Nonparametric sparse hierarchical models describe V1 fMRI responses to natural images
Vincent Q. Vu, Bin Yu, Thomas Naselaris, Kendrick Kay, Jack Gallant, Pradeep Ravikumar
We propose a novel hierarchical, nonlinear model that predicts brain activity in area V1 evoked by natural images. In the study reported here brain activity was measured by means of functional magnetic resonance imaging (fMRI), a noninvasive technique that provides an indirect measure of neural activity pooled over a small volume (~ 2mm cube) of brain tissue. Our model, which we call the SpAM V1 model, is based on the reasonable assumption that fMRI measurements reflect the (possibly nonlinearly) pooled, rectified output of a large population of simple and complex cells in V1. It has a hierarchical filtering stage that consists of three layers: model simple cells, model complex cells, and a third layer in which the complex cells are linearly pooled (called â■■pooled-complexâ■■ cells). The pooling stage then obtains the measured fMRI signals as a sparse additive model (SpAM) in which a sparse nonparametric (nonlinear) combination of model complex cell and model pooled-complex cell outputs are summed. Our results show that the SpAM V1 model predicts fMRI responses evoked by natural images better than a benchmark model that only provides linear pooling of model complex cells. Furthermore, the spatial receptive fields, frequency tuning and orientation tuning curves of the SpAM V1 model estimated for each voxel appears to be consistent with the known properties of V1, and with previous analyses of this data set. A visualization procedure applied to the SpAM V1 model shows that most of the nonlinear pooling consists of simple compressive or saturating nonlinearities.
****

A Scalable Hierarchical Distributed Language Model
Andriy Mnih, Geoffrey E. Hinton
Neural probabilistic language models (NPLMs) have been shown to be competitive with and occasionally superior to the widely-used n-gram language models. The main drawback of NPLMs is their extremely long training and testing times. Morin and Bengio have proposed a hierarchical language model built around a binary tree of words that was two orders of magnitude faster than the non-hierarchical language model it was based on. However, it performed considerably worse than its non-hierarchical counterpart in spite of using a word tree created using expert knowledge. We introduce a fast hierarchical language model along with a simple feature-based algorithm for automatic construction of word trees from the data. We then show that the resulting models can outperform non-hierarchical models and achieve state-of-the-art performance.
****

Supervised Exponential Family Principal Component Analysis via Convex Optimization
Yuhong Guo
Recently, supervised dimensionality reduction has been gaining attention, owing to the realization that data labels are often available and strongly suggest important underlying structures in the data. In this paper, we present a novel convex supervised dimensionality reduction approach based on exponential family PCA and provide a simple but novel form to project new testing data into the embedded space. This convex approach successfully avoids the local optima of the EM learning. Moreover, by introducing a sample-based multinomial approximation to exponential family models, it avoids the limitation of the prevailing Gaussian assumptions of standard PCA, and produces a kernelized formulation for nonlinear supervised dimensionality reduction. A training algorithm is then devised based on a subgradient bundle method, whose scalability can be gained through a coordinate descent procedure. The advantage of our global optimization approach is demonstrated by empirical results over both synthetic and real data.
****

Stochastic Relational Models for Large-scale Dyadic Data using MCMC

Shenghuo Zhu, Kai Yu, Yihong Gong
Stochastic relational models provide a rich family of choices for learning and p
redicting dyadic data between two sets of entities. It generalizes matrix factor
ization to a supervised learning problem that utilizes attributes of objects in
a hierarchical Bayesian framework. Previously empirical Bayesian inference was a
pplied, which is however not scalable when the size of either object sets become
s tens of thousands. In this paper, we introduce a Markov chain Monte Carlo (MCM
C) algorithm to scale the model to very large-scale dyadic data. Both superior s
calability and predictive accuracy are demonstrated on a collaborative filtering
 problem, which involves tens of thousands users and a half million items.
****

Online Models for Content Optimization
Deepak Agarwal, Bee-chung Chen, Pradheep Elango, Nitin Motgi, Seung-taek Park, R
aghu Ramakrishnan, Scott Roy, Joe Zachariah
We describe a new content publishing system that selects articles to serve to a
user, choosing from an editorially programmed pool that is frequently refreshed.
 It is now deployed on a major Internet portal, and selects articles to serve to
 hundreds of millions of user visits per day, significantly increasing the numbe
r of user clicks over the original manual approach, in which editors periodicall
y selected articles to display. Some of the challenges we face include a dynamic
 content pool, short article lifetimes, non-stationary click-through rates, and
extremely high traffic volumes. The fundamental problem we must solve is to quic
kly identify which items are popular(perhaps within different user segments), an
d to exploit them while they remain current. We must also explore the underlying
 pool constantly to identify promising alternatives, quickly discarding poor per
formers. Our approach is based on tracking per article performance in near real
time through online models. We describe the characteristics and constraints of o
ur application setting, discuss our design choices, and show the importance and
effectiveness of coupling online models with a simple randomization procedure. W
e discuss the challenges encountered in a production online content-publishing e
nvironment and highlight issues that deserve careful attention. Our analysis of
this application also suggests a number of future research avenues.
****

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

Hierarchical Fisher Kernels for Longitudinal Data
Zhengdong Lu, Jeffrey Kaye, Todd Leen
We develop new techniques for time series classification based on hierarchical B
ayesian generative models (called mixed-effect models) and the Fisher kernel der
ived from them. A key advantage of the new formulation is that one can compute t
he Fisher information matrix despite varying sequence lengths and sampling times
. We therefore can avoid the ad hoc replacement of Fisher information matrix wit
h the identity matrix commonly used in literature, which destroys the geometrica
l grounding of the kernel construction. In contrast, our construction retains th
e proper geometric structure resulting in a kernel that is properly invariant un
der change of coordinates in the model parameter space. Experiments on detecting
 cognitive decline show that classifiers based on the proposed kernel out-perfor
m those based on generative models and other feature extraction routines.
****

Correlated Bigram LSA for Unsupervised Language Model Adaptation
Yik-cheung Tam, Tanja Schultz
We propose using correlated bigram LSA for unsupervised LM adaptation for automa
tic speech recognition. The model is trained using efficient variational EM and
smoothed using the proposed fractional Kneser-Ney smoothing which handles fracti
onal counts. Our approach can be scalable to large training corpora via bootstra

pping of bigram LSA from unigram LSA. For LM adaptation, unigram and bigram LSA are integrated into the background N-gram LM via marginal adaptation and linear interpolation respectively. Experimental results show that applying unigram and bigram LSA together yields 6%--8% relative perplexity reduction and 0.6% absolute character error rates (CER) reduction compared to applying only unigram LSA on the Mandarin RT04 test set. Comparing with the unadapted baseline, our approach reduces the absolute CER by 1.2%.
****

The Infinite Factorial Hidden Markov Model
Jurgen Gael, Yee Teh, Zoubin Ghahramani
We introduces a new probability distribution over a potentially infinite number of binary Markov chains which we call the Markov Indian buffet process. This process extends the IBP to allow temporal dependencies in the hidden variables. We use this stochastic process to build a nonparametric extension of the factorial hidden Markov model. After working out an inference scheme which combines slice sampling and dynamic programming we demonstrate how the infinite factorial hidden Markov model can be used for blind source separation.
****

Sparse Signal Recovery Using Markov Random Fields
Volkan Cevher, Marco Duarte, Chinmay Hegde, Richard Baraniuk
Compressive Sensing (CS) combines sampling and compression into a single sub-Nyquist linear measurement process for sparse and compressible signals. In this paper, we extend the theory of CS to include signals that are concisely represented in terms of a graphical model. In particular, we use Markov Random Fields (MRFs) to represent sparse signals whose nonzero coefficients are clustered. Our new model-based reconstruction algorithm, dubbed Lattice Matching Pursuit (LaMP), stably recovers MRF-modeled signals using many fewer measurements and computations than the current state-of-the-art algorithms.
****

Clustering via LP-based Stabilities
Nikos Komodakis, Nikos Paragios, Georgios Tziritas
A novel center-based clustering algorithm is proposed in this paper. We first formulate clustering as an NP-hard linear integer program and we then use linear programming and the duality theory to derive the solution of this optimization problem. This leads to an efficient and very general algorithm, which works in the dual domain, and can cluster data based on an arbitrary set of distances. Despite its generality, it is independent of initialization (unlike EM-like methods such as K-means), has guaranteed convergence, and can also provide online optimality bounds about the quality of the estimated clustering solutions. To deal with the most critical issue in a center-based clustering algorithm (selection of cluster centers), we also introduce the notion of stability of a cluster center, which is a well defined LP-based quantity that plays a key role to our algorithm's success. Furthermore, we also introduce, what we call, the margins (another key ingredient in our algorithm), which can be roughly thought of as dual counterparts to stabilities and allow us to obtain computationally efficient approximations to the latter. Promising experimental results demonstrate the potentials of our method.
****

Spike Feature Extraction Using Informative Samples
Zhi Yang, Qi Zhao, Wentai Liu
This paper presents a spike feature extraction algorithm that targets real-time spike sorting and facilitates miniaturized microchip implementation. The proposed algorithm has been evaluated on synthesized waveforms and experimentally recorded sequences. When compared with many spike sorting approaches our algorithm demonstrates improved speed, accuracy and allows unsupervised execution. A preliminary hardware implementation has been realized using an integrated microchip interfaced with a personal computer.
****

Fast Computation of Posterior Mode in Multi-Level Hierarchical Models
Liang Zhang, Deepak Agarwal

Multi-level hierarchical models provide an attractive framework for incorporatin
g correlations induced in a response variable organized in a hierarchy. Model fi
tting is challenging, especially for hierarchies with large number of nodes. We
provide a novel algorithm based on a multi-scale Kalman filter that is both scal
able and easy to implement. For non-Gaussian responses, quadratic approximation
to the log-likelihood results in biased estimates. We suggest a bootstrap strate
gy to correct such biases. Our method is illustrated through simulation studies
and analyses of real world data sets in health care and online advertising.
****

An improved estimator of Variance Explained in the presence of noise
Ralf Haefner, Bruce Cumming
A crucial part of developing mathematical models of how the brain works is the q
uantification of their success. One of the most widely-used metrics yields the p
ercentage of the variance in the data that is explained by the model. Unfortunat
ely, this metric is biased due to the intrinsic variability in the data. This va
riability is in principle unexplainable by the model. We derive a simple analyti
cal modification of the traditional formula that significantly improves its accu
racy (as measured by bias) with similar or better precision (as measured by mean
-square error) in estimating the true underlying Variance Explained by the model
 class. Our estimator advances on previous work by a) accounting for the uncerta
inty in the noise estimate, b) accounting for overfitting due to free model para
meters mitigating the need for a separate validation data set and c) adding a co
nditioning term. We apply our new estimator to binocular disparity tuning curves
 of a set of macaque V1 neurons and find that on a population level almost all o
f the variance unexplained by Gabor functions is attributable to noise.
****

The Infinite Hierarchical Factor Regression Model
Piyush Rai, Hal Daume
We propose a nonparametric Bayesian factor regression model that accounts for un
certainty in the number of factors, and the relationship between factors. To acc
omplish this, we propose a sparse variant of the Indian Buffet Process and coupl
e this with a hierarchical model over factors, based on Kingman's coalescent. We
 apply this model to two problems (factor analysis and factor regression) in gen
e-expression data analysis.
****

Deep Learning with Kernel Regularization for Visual Recognition
Kai Yu, Wei Xu, Yihong Gong
In this paper we focus on training deep neural networks for visual recognition t
asks. One challenge is the lack of an informative regularization on the network
parameters, to imply a meaningful control on the computed function. We propose a
 training strategy that takes advantage of kernel methods, where an existing ker
nel function represents useful prior knowledge about the learning task of intere
st. We derive an efficient algorithm using stochastic gradient descent, and demo
nstrate very positive results in a wide range of visual recognition tasks.
****

Learning Transformational Invariants from Natural Movies
Charles Cadieu, Bruno Olshausen
We describe a hierarchical, probabilistic model that learns to extract complex m
otion from movies of the natural environment. The model consists of two hidden l
ayers: the first layer produces a sparse representation of the image that is exp
ressed in terms of local amplitude and phase variables. The second layer learns
the higher-order structure among the time-varying phase variables. After trainin
g on natural movies, the top layer units discover the structure of phase-shifts
within the first layer. We show that the top layer units encode transformational
 invariants: they are selective for the speed and direction of a moving pattern,
 but are invariant to its spatial structure (orientation/spatial-frequency). The
 diversity of units in both the intermediate and top layers of the model provide
s a set of testable predictions for representations that might be found in V1 an
d MT. In addition, the model demonstrates how feedback from higher levels can in
fluence representations at lower levels as a by-product of inference in a graphi

cal model.
****

On Bootstrapping the ROC Curve
Patrice Bertail, Stéphan Clémençcon, Nicolas Vayatis
This paper is devoted to thoroughly investigating how to bootstrap the ROC curve , a widely used visual tool for evaluating the accuracy of test/scoring statisti cs in the bipartite setup. The issue of confidence bands for the ROC curve is co nsidered and a resampling procedure based on a smooth version of the empirical d istribution called the smoothed bootstrap" is introduced. Theoretical arguments and simulation results are presented to show that the "smoothed bootstrap" is pr eferable to a "naive" bootstrap in order to construct accurate confidence bands. "
****

QUIC-SVD: Fast SVD Using Cosine Trees
Michael Holmes, Jr. Isbell, Charles Lee, Alexander Gray
The Singular Value Decomposition is a key operation in many machine learning met hods. Its computational cost, however, makes it unscalable and impractical for t he massive-sized datasets becoming common in applications. We present a new meth od, QUIC-SVD, for fast approximation of the full SVD with automatic sample size minimization and empirical relative error control. Previous Monte Carlo approach es have not addressed the full SVD nor benefited from the efficiency of automati c, empirically-driven sample sizing. Our empirical tests show speedups of severa l orders of magnitude over exact SVD. Such scalability should enable QUIC-SVD to meet the needs of a wide array of methods and applications.
****

A Massively Parallel Digital Learning Processor
Hans Graf, Srihari Cadambi, Venkata Jakkula, Murugan Sankaradass, Eric Cosatto, Srimat Chakradhar, Igor Dourdanovic
We present a new, massively parallel architecture for accelerating machine learn ing algorithms, based on arrays of variable-resolution arithmetic vector process ing elements (VPE). Groups of VPEs operate in SIMD (single instruction multiple data) mode, and each group is connected to an independent memory bank. In this w ay memory bandwidth scales with the number of VPE, and the main data flows are l ocal, keeping power dissipation low. With 256 VPEs, implemented on two FPGA (fie ld programmable gate array) chips, we obtain a sustained speed of 19 GMACS (bill ion multiply-accumulate per sec.) for SVM training, and 86 GMACS for SVM classif ication. This performance is more than an order of magnitude higher than that of any FPGA implementation reported so far. The speed on one FPGA is similar to th e fastest speeds published on a Graphics Processor for the MNIST problem, despit e a clock rate of the FPGA that is six times lower. High performance at low cloc k rates makes this massively parallel architecture particularly attractive for e mbedded applications, where low power dissipation is critical. Tests with Convol utional Neural Networks and other learning algorithms are under way now.
****

Characterizing response behavior in multisensory perception with conflicting cue s
Rama Natarajan, Iain Murray, Ladan Shams, Richard Zemel
We explore a recently proposed mixture model approach to understand- ing interac tions between con■icting sensory cues. Alternative model for- mulations, differi ng in their sensory noise models and inference methods, are compared based on th eir ■t to experimental data. Heavy-tailed sen- sory likelihoods yield a better d escription of the subjects’ response behavior than standard Gaussian noise model s. We study the underlying cause for this result, and then present several testa ble predictions of these models.
****

The Gaussian Process Density Sampler
Iain Murray, David MacKay, Ryan P. Adams
We present the Gaussian Process Density Sampler (GPDS), an exchangeable generati ve model for use in nonparametric Bayesian density estimation. Samples drawn fro m the GPDS are consistent with exact, independent samples from a fixed density f

unction that is a transformation of a function drawn from a Gaussian process pri
or. Our formulation allows us to infer an unknown density from data using Markov
 chain Monte Carlo, which gives samples from the posterior distribution over den
sity functions and from the predictive distribution on data space. We can also i
nfer the hyperparameters of the Gaussian process. We compare this density modeli
ng technique to several existing techniques on a toy problem and a skull-reconst
ruction task.
****

Cyclizing Clusters via Zeta Function of a Graph
Deli Zhao, Xiaoou Tang
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.
****

Integrating Locally Learned Causal Structures with Overlapping Variables
David Danks, Clark Glymour, Robert Tillman
In many domains, data are distributed among datasets that share only some variab
les; other recorded variables may occur in only one dataset. There are several a
symptotically correct, informative algorithms that search for causal information
 given a single dataset, even with missing values and hidden variables. There ar
e, however, no such reliable procedures for distributed data with overlapping va
riables, and only a single heuristic procedure (Structural EM). This paper descr
ibes an asymptotically correct procedure, ION, that provides all the information
 about structure obtainable from the marginal independence relations. Using simu
lated and real data, the accuracy of ION is compared with that of Structural EM,
 and with inference on complete, unified data.
****

A mixture model for the evolution of gene expression in non-homogeneous datasets
Gerald Quon, Yee Teh, Esther Chan, Timothy Hughes, Michael Brudno, Quaid Morris
We address the challenge of assessing conservation of gene expression in complex
, non-homogeneous datasets. Recent studies have demonstrated the success of prob
abilistic models in studying the evolution of gene expression in simple eukaryot
ic organisms such as yeast, for which measurements are typically scalar and inde
pendent. Models capable of studying expression evolution in much more complex or
ganisms such as vertebrates are particularly important given the medical and sci
entific interest in species such as human and mouse. We present a statistical mo
del that makes a number of significant extensions to previous models to enable c
haracterization of changes in expression among highly complex organisms. We demo
nstrate the efficacy of our method on a microarray dataset containing diverse ti
ssues from multiple vertebrate species. We anticipate that the model will be inv
aluable in the study of gene expression patterns in other diverse organisms as w
ell, such as worms and insects.
****

An Homotopy Algorithm for the Lasso with Online Observations
Pierre Garrigues, Laurent Ghaoui
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.
****

Adaptive Martingale Boosting
Phil Long, Rocco Servedio
In recent work Long and Servedio LS05short presented a ``martingale boosting'' a
lgorithm that works by constructing a branching program over weak classifiers an
d has a simple analysis based on elementary properties of random walks. LS05shor
t showed that this martingale booster can tolerate random classification noise w
hen it is run with a noise-tolerant weak learner; however, a drawback of the alg
orithm is that it is not adaptive, i.e. it cannot effectively take advantage of
variation in the quality of the weak classifiers it receives. In this paper we p

resent a variant of the original martingale boosting algorithm and prove that it is adaptive. This adaptiveness is achieved by modifying the original algorithm so that the random walks that arise in its analysis have different step size depending on the quality of the weak learner at each stage. The new algorithm inherits the desirable properties of the original LS05short algorithm, such as random classification noise tolerance, and has several other advantages besides adaptiveness: it requires polynomially fewer calls to the weak learner than the original algorithm, and it can be used with confidence-rated weak hypotheses that output real values rather than Boolean predictions.

****

Fast High-dimensional Kernel Summations Using the Monte Carlo Multipole Method

Dongryeol Lee, Alexander Gray

We propose a new fast Gaussian summation algorithm for high-dimensional datasets with high accuracy. First, we extend the original fast multipole-type methods to use approximation schemes with both hard and probabilistic error. Second, we utilize a new data structure called subspace tree which maps each data point in the node to its lower dimensional mapping as determined by any linear dimension reduction method such as PCA. This new data structure is suitable for reducing the cost of each pairwise distance computation, the most dominant cost in many kernel methods. Our algorithm guarantees probabilistic relative error on each kernel sum, and can be applied to high-dimensional Gaussian summations which are ubiquitous inside many kernel methods as the key computational bottleneck. We provide empirical speedup results on low to high-dimensional datasets up to 89 dimensions.

****

ICA based on a Smooth Estimation of the Differential Entropy

Lev Faivishevsky, Jacob Goldberger

In this paper we introduce the MeanNN approach for estimation of main information theoretic measures such as differential entropy, mutual information and divergence. As opposed to other nonparametric approaches the MeanNN results in smooth differentiable functions of the data samples with clear geometrical interpretation. Then we apply the proposed estimators to the ICA problem and obtain a smooth expression for the mutual information that can be analytically optimized by gradient descent methods. The improved performance on the proposed ICA algorithm is demonstrated on standard tests in comparison with state-of-the-art techniques.

****

Support Vector Machines with a Reject Option

Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, Stéphane Canu

We consider the problem of binary classification where the classifier may abstain instead of classifying each observation. The Bayes decision rule for this setup, known as Chow's rule, is defined by two thresholds on posterior probabilities. From simple desiderata, namely the consistency and the sparsity of the classifier, we derive the double hinge loss function that focuses on estimating conditional probabilities only in the vicinity of the threshold points of the optimal decision rule. We show that, for suitable kernel machines, our approach is universally consistent. We cast the problem of minimizing the double hinge loss as a quadratic program akin to the standard SVM optimization problem and propose an active set method to solve it efficiently. We finally provide preliminary experimental results illustrating the interest of our constructive approach to devising loss functions.

****

Generative versus discriminative training of RBMs for classification of fMRI images

Tanya Schmah, Geoffrey E. Hinton, Steven Small, Stephen Strother, Richard Zemel

Neuroimaging datasets often have a very large number of voxels and a very small number of training cases, which means that overfitting of models for this data can become a very serious problem. Working with a set of fMRI images from a study on stroke recovery, we consider a classification task for which logistic regression performs poorly, even when L1- or L2- regularized. We show that much better discrimination can be achieved by fitting a generative model to each separate c

ondition and then seeing which model is most likely to have generated the data. We compare discriminative training of exactly the same set of models, and we also consider convex blends of generative and discriminative training.
****

Particle Filter-based Policy Gradient in POMDPs
Pierre-arnaud Coquelin, Romain Deguest, Rémi Munos
Our setting is a Partially Observable Markov Decision Process with continuous state, observation and action spaces. Decisions are based on a Particle Filter for estimating the belief state given past observations. We consider a policy gradient approach for parameterized policy optimization. For that purpose, we investigate sensitivity analysis of the performance measure with respect to the parameters of the policy, focusing on Finite Difference (FD) techniques. We show that the naive FD is subject to variance explosion because of the non-smoothness of the resampling procedure. We propose a more sophisticated FD method which overcomes this problem and establish its consistency.
****

Algorithms for Infinitely Many-Armed Bandits
Yizao Wang, Jean-yves Audibert, Rémi Munos
We consider multi-armed bandit problems where the number of arms is larger than the possible number of experiments. We make a stochastic assumption on the mean-reward of a new selected arm which characterizes its probability of being a near-optimal arm. Our assumption is weaker than in previous works. We describe algorithms based on upper-confidence-bounds applied to a restricted set of randomly selected arms and provide upper-bounds on the resulting expected regret. We also derive a lower-bound which matchs (up to logarithmic factors) the upper-bound in some cases.
****

On the asymptotic equivalence between differential Hebbian and temporal difference learning using a local third factor
Christoph Kolodziejski, Bernd Porr, Minija Tamosiunaite, Florentin Wörgötter
In this theoretical contribution we provide mathematical proof that two of the most important classes of network learning - correlation-based differential Hebbian learning and reward-based temporal difference learning - are asymptotically equivalent when timing the learning with a local modulatory signal. This opens the opportunity to consistently reformulate most of the abstract reinforcement learning framework from a correlation based perspective that is more closely related to the biophysics of neurons.
****

Gates
Tom Minka, John Winn
Gates are a new notation for representing mixture models and context-sensitive independence in factor graphs. Factor graphs provide a natural representation for message-passing algorithms, such as expectation propagation. However, message passing in mixture models is not well captured by factor graphs unless the entire mixture is represented by one factor, because the message equations have a containment structure. Gates capture this containment structure graphically, allowing both the independences and the message-passing equations for a model to be readily visualized. Different variational approximations for mixture models can be understood as different ways of drawing the gates in a model. We present general equations for expectation propagation and variational message passing in the presence of gates.
****

Multi-Level Active Prediction of Useful Image Annotations for Recognition
Sudheendra Vijayanarasimhan, Kristen Grauman
We introduce a framework for actively learning visual categories from a mixture of weakly and strongly labeled image examples. We propose to allow the category-learner to strategically choose what annotations it receives---based on both the expected reduction in uncertainty as well as the relative costs of obtaining each annotation. We construct a multiple-instance discriminative classifier based on the initial training data. Then all remaining unlabeled and weakly labeled ex

amples are surveyed to actively determine which annotation ought to be requested next. After each request, the current classifier is incrementally updated. Unlike previous work, our approach accounts for the fact that the optimal use of manual annotation may call for a combination of labels at multiple levels of granularity (e.g., a full segmentation on some images and a present/absent flag on others). As a result, it is possible to learn more accurate category models with a lower total expenditure of manual annotation effort.
****

MAS: a multiplicative approximation scheme for probabilistic inference
Ydo Wexler, Christopher Meek
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Learning Hybrid Models for Image Annotation with Partially Labeled Data
Xuming He, Richard Zemel
Extensive labeled data for image annotation systems, which learn to assign class labels to image regions, is difficult to obtain. We explore a hybrid model framework for utilizing partially labeled data that integrates a generative topic model for image appearance with discriminative label prediction. We propose three alternative formulations for imposing a spatial smoothness prior on the image labels. Tests of the new models and some baseline approaches on two real image datasets demonstrate the effectiveness of incorporating the latent structure.
****

Efficient Sampling for Gaussian Process Inference using Control Variables
Neil Lawrence, Magnus Rattray, Michalis Titsias
Sampling functions in Gaussian process (GP) models is challenging because of the highly correlated posterior distribution. We describe an efficient Markov chain Monte Carlo algorithm for sampling from the posterior process of the GP model. This algorithm uses control variables which are auxiliary function values that provide a low dimensional representation of the function. At each iteration, the algorithm proposes new values for the control variables and generates the function from the conditional GP prior. The control variable input locations are found by continuously minimizing an objective function. We demonstrate the algorithm on regression and classification problems and we use it to estimate the parameters of a differential equation model of gene regulation.
****

An Online Algorithm for Maximizing Submodular Functions
Matthew Streeter, Daniel Golovin
We present an algorithm for solving a broad class of online resource allocation problems. Our online algorithm can be applied in environments where abstract jobs arrive one at a time, and one can complete the jobs by investing time in a number of abstract activities, according to some schedule. We assume that the fraction of jobs completed by a schedule is a monotone, submodular function of a set of pairs (v,t), where t is the time invested in activity v. Under this assumption, our online algorithm performs near-optimally according to two natural metrics: (i) the fraction of jobs completed within time T, for some fixed deadline T > 0, and (ii) the average time required to complete each job. We evaluate our algorithm experimentally by using it to learn, online, a schedule for allocating CPU time among solvers entered in the 2007 SAT solver competition.
****

On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization
Sham M. Kakade, Karthik Sridharan, Ambuj Tewari
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Bayesian Exponential Family PCA
Shakir Mohamed, Zoubin Ghahramani, Katherine A. Heller
Principal Components Analysis (PCA) has become established as one of the key tools for dimensionality reduction when dealing with real valued data. Approaches such as exponential family PCA and non-negative matrix factorisation have successfully extended PCA to non-Gaussian data types, but these techniques fail to take advantage of Bayesian inference and can suffer from problems of overfitting and poor generalisation. This paper presents a fully probabilistic approach to PCA, which is generalised to the exponential family, based on Hybrid Monte Carlo sampling. We describe the model which is based on a factorisation of the observed data matrix, and show performance of the model on both synthetic and real data.
****
Sequential effects: Superstition or rational behavior?
Angela J. Yu, Jonathan D. Cohen
In a variety of behavioral tasks, subjects exhibit an automatic and apparently sub-optimal sequential effect: they respond more rapidly and accurately to a stimulus if it reinforces a local pattern in stimulus history, such as a string of repetitions or alternations, compared to when it violates such a pattern. This is often the case even if the local trends arise by chance in the context of a randomized design, such that stimulus history has no predictive power. In this work, we use a normative Bayesian framework to examine the hypothesis that such idiosyncrasies may reflect the inadvertent engagement of fundamental mechanisms critical for adapting to changing statistics in the natural environment. We show that prior belief in non-stationarity can induce experimentally observed sequential effects in an otherwise Bayes-optimal algorithm. The Bayesian algorithm is shown to be well approximated by linear-exponential filtering of past observations, a feature also apparent in the behavioral data. We derive an explicit relationship between the parameters and computations of the exact Bayesian algorithm and those of the approximate linear-exponential filter. Since the latter is equivalent to a leaky-integration process, a commonly used model of neuronal dynamics underlying perceptual decision-making and trial-to-trial dependencies, our model provides a principled account of why such dynamics are useful. We also show that near-optimal tuning of the leaky-integration process is possible, using stochastic gradient descent based only on the noisy binary inputs. This is a proof of concept that not only can neurons implement near-optimal prediction based on standard neuronal dynamics, but that they can also learn to tune the processing parameters without explicitly representing probabilities.
****
PSDBoost: Matrix-Generation Linear Programming for Positive Semidefinite Matrices Learning
Chunhua Shen, Alan Welsh, Lei Wang
In this work, we consider the problem of learning a positive semidefinite matrix. The critical issue is how to preserve positive semidefiniteness during the course of learning. Our algorithm is mainly inspired by LPBoost [1] and the general greedy convex optimization framework of Zhang [2]. We demonstrate the essence of the algorithm, termed PSDBoost (positive semidefinite Boosting), by focusing on a few different applications in machine learning. The proposed PSDBoost algorithm extends traditional Boosting algorithms in that its parameter is a positive semidefinite matrix with trace being one instead of a classifier. PSDBoost is based on the observation that any trace-one positive semidefinitematrix can be decomposed into linear convex combinations of trace-one rank-one matrices, which serve as base learners of PSDBoost. Numerical experiments are presented.
****
Kernel-ARMA for Hand Tracking and Brain-Machine interfacing During 3D Motor Control
Lavi Shpigelman, Hagai Lalazar, Eilon Vaadia
Using machine learning algorithms to decode intended behavior from neural activity serves a dual purpose. First, these tools can be used to allow patients to interact with their environment through a Brain-Machine Interface (BMI). Second, analysis of the characteristics of such methods can reveal the significance of va

rious features of neural activity, stimuli and responses to the encoding-decodin g task. In this study we adapted, implemented and tested a machine learning meth od, called Kernel Auto-Regressive Moving Average (KARMA), for the task of inferr ing movements from neural activity in primary motor cortex. Our version of this algorithm is used in an on-line learning setting and is updated when feedback fr om the last inferred sequence become available. We first used it to track real h and movements executed by a monkey in a standard 3D motor control task. We then applied it in a closed-loop BMI setting to infer intended movement, while arms w ere restrained, allowing a monkey to perform the task using the BMI alone. KARMA is a recurrent method that learns a nonlinear model of output dynamics. It uses similarity functions (termed kernels) to compare between inputs. These kernels can be structured to incorporate domain knowledge into the method. We compare KA RMA to various state-of-the-art methods by evaluating tracking performance and p resent results from the KARMA based BMI experiments.
****

Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of \b oldmathℓ_1-regularized MLE
Garvesh Raskutti, Bin Yu, Martin J. Wainwright, Pradeep Ravikumar
We consider the problem of estimating the graph structure associated with a Gaus sian Markov random field (GMRF) from i.i.d. samples. We study the performance of study the performance of the 1 -regularized maximum likelihood estimator in the high-dimensional setting, where the number of nodes in the graph p, the number of edges in the graph s and the maximum node degree d, are allowed to grow as a function of the number of samples n. Our main result provides sufficient conditi ons on (n, p, d) for the 1 -regularized MLE estimator to recover all the edges o f the graph with high probability. Under some conditions on the model covariance , we show that model selection can be achieved for sample sizes n = (d2 log(p)), with the error decaying as O(exp(-c log(p))) for some constant c. We illustrate our theoretical results via simulations and show good correspondences between t he theoretical predictions and behavior in simulations.
****

Stress, noradrenaline, and realistic prediction of mouse behaviour using reinfor cement learning
Carmen Sandi, Wulfram Gerstner, Gediminas Lukšys
Suppose we train an animal in a conditioning experiment. Can one predict how a g iven animal, under given experimental conditions, would perform the task? Since various factors such as stress, motivation, genetic background, and previous err ors in task performance can influence animal behaviour, this appears to be a ver y challenging aim. Reinforcement learning (RL) models have been successful in mo deling animal (and human) behaviour, but their success has been limited because of uncertainty as to how to set meta-parameters (such as learning rate, exploita tion-exploration balance and future reward discount factor) that strongly influe nce model performance. We show that a simple RL model whose metaparameters are c ontrolled by an artificial neural network, fed with inputs such as stress, affec tive phenotype, previous task performance, and even neuromodulatory manipulation s, can successfully predict mouse behaviour in the "hole-box" - a simple conditi oning task. Our results also provide important insights on how stress and anxiet y affect animal learning, performance accuracy, and discounting of future reward s, and on how noradrenergic systems can interact with these processes.
****

How memory biases affect information transmission: A rational analysis of serial reproduction
Jing Xu, Thomas Griffiths
Many human interactions involve pieces of information being passed from one pers on to another, raising the question of how this process of information transmiss ion is affected by the capacities of the agents involved. In the 1930s, Sir Fred eric Bartlett explored the influence of memory biases in â■■serial reproductionâ ■■ of information, in which one personâ■■s reconstruction of a stimulus from mem ory becomes the stimulus seen by the next person. These experiments were done us ing relatively uncontrolled stimuli such as pictures and stories, but suggested

that serial reproduction would transform information in a way that reflected the biases inherent in memory. We formally analyze serial reproduction using a Bayesian model of reconstruction from memory, giving a general result characterizing the effect of memory biases on information transmission. We then test the predictions of this account in two experiments using simple one-dimensional stimuli. Our results provide theoretical and empirical justification for the idea that serial reproduction reflects memory biases.
****

Diffeomorphic Dimensionality Reduction
Christian Walder, Bernhard Schölkopf
This paper introduces a new approach to constructing meaningful lower dimensional representations of sets of data points. We argue that constraining the mapping between the high and low dimensional spaces to be a diffeomorphism is a natural way of ensuring that pairwise distances are approximately preserved. Accordingly we develop an algorithm which diffeomorphically maps the data near to a lower dimensional subspace and then projects onto that subspace. The problem of solving for the mapping is transformed into one of solving for an Eulerian flow field which we compute using ideas from kernel methods. We demonstrate the efficacy of our approach on various real world data sets.
****

Using Bayesian Dynamical Systems for Motion Template Libraries
Silvia Chiappa, Jens Kober, Jan Peters
Motor primitives or motion templates have become an important concept for both modeling human motor control as well as generating robot behaviors using imitation learning. Recent impressive results range from humanoid robot movement generation to timing models of human motions. The automatic generation of skill libraries containing multiple motion templates is an important step in robot learning. Such a skill learning system needs to cluster similar movements together and represent each resulting motion template as a generative model which is subsequently used for the execution of the behavior by a robot system. In this paper, we show how human trajectories captured as multidimensional time-series can be clustered using Bayesian mixtures of linear Gaussian state-space models based on the similarity of their dynamics. The appropriate number of templates is automatically determined by enforcing a parsimonious parametrization. As the resulting model is intractable, we introduce a novel approximation method based on variational Bayes, which is especially designed to enable the use of efficient inference algorithms. On recorded human Balero movements, this method is not only capable of finding reasonable motion templates but also yields a generative model which works well in the execution of this complex task on a simulated anthropomorphic SARCOS arm.
****

An Efficient Sequential Monte Carlo Algorithm for Coalescent Clustering
Dilan Gorur, Yee Teh
We propose an efficient sequential Monte Carlo inference scheme for the recently proposed coalescent clustering model (Teh et al, 2008). Our algorithm has a quadratic runtime while those in (Teh et al, 2008) is cubic. In experiments, we were surprised to find that in addition to being more efficient, it is also a better sequential Monte Carlo sampler than the best in (Teh et al, 2008), when measured in terms of variance of estimated likelihood and effective sample size.
****

Bounding Performance Loss in Approximate MDP Homomorphisms
Jonathan Taylor, Doina Precup, Prakash Panagaden
We define a metric for measuring behavior similarity between states in a Markov decision process (MDP), in which action similarity is taken into account. We show that the kernel of our metric corresponds exactly to the classes of states defined by MDP homomorphisms (Ravindran \& Barto, 2003). We prove that the difference in the optimal value function of different states can be upper-bounded by the value of this metric, and that the bound is tighter than that provided by bisimulation metrics (Ferns et al. 2004, 2005). Our results hold both for discrete and for continuous actions. We provide an algorithm for constructing approximate h

omomorphisms, by using this metric to identify states that can be grouped togeth
er, as well as actions that can be matched. Previous research on this topic is b
ased mainly on heuristics.
****

Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks
Alex Graves, Jürgen Schmidhuber
Offline handwriting recognition---the transcription of images of handwritten tex
t---is an interesting task, in that it combines computer vision with sequence le
arning. In most systems the two elements are handled separately, with sophistica
ted preprocessing techniques used to extract the image features and sequential m
odels such as HMMs used to provide the transcriptions. By combining two recent i
nnovations in neural networks---multidimensional recurrent neural networks and c
onnectionist temporal classification---this paper introduces a globally trained
offline handwriting recogniser that takes raw pixel data as input. Unlike compet
ing systems, it does not require any alphabet specific preprocessing, and can th
erefore be used unchanged for any language. Evidence of its generality and power
 is provided by data from a recent international Arabic recognition competition,
 where it outperformed all entries (91.4% accuracy compared to 87.2% for the com
petition winner) despite the fact that neither author understands a word of Arab
ic.
****

Robust Near-Isometric Matching via Structured Learning of Graphical Models
Alex Smola, Julian Mcauley, Tibério Caetano
Models for near-rigid shape matching are typically based on distance-related fea
tures, in order to infer matches that are consistent with the isometric assumpti
on. However, real shapes from image datasets, even when expected to be related b
y almost isometric" transformations, are actually subject not only to noise but
also, to some limited degree, to variations in appearance and scale. In this pap
er, we introduce a graphical model that parameterises appearance, distance, and
angle features and we learn all of the involved parameters via structured predic
tion. The outcome is a model for near-rigid shape matching which is robust in th
e sense that it is able to capture the possibly limited but still important scal
e and appearance variations. Our experimental results reveal substantial improve
ments upon recent successful models, while maintaining similar running times."
****

Fast Prediction on a Tree
Mark Herbster, Massimiliano Pontil, Sergio Galeano
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.
****

Exact Convex Confidence-Weighted Learning
Koby Crammer, Mark Dredze, Fernando Pereira
Confidence-weighted (CW) learning [6], an online learning method for linear clas
sifiers, maintains a Gaussian distributions over weight vectors, with a covarian
ce matrix that represents uncertainty about weights and correlations. Confidence
 constraints ensure that a weight vector drawn from the hypothesis distribution
correctly classifies examples with a specified probability. Within this framewor
k, we derive a new convex form of the constraint and analyze it in the mistake b
ound model. Empirical evaluation with both synthetic and text data shows our ver
sion of CW learning achieves lower cumulative and out-of-sample errors than comm
only used first-order and second-order online methods.
****

The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Red
undancy Reduction
Fabian Sinz, Matthias Bethge
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.
****

Regularized Co-Clustering with Dual Supervision
Vikas Sindhwani, Jianying Hu, Aleksandra Mojsilovic
By attempting to simultaneously partition both the rows (examples) and columns (features) of a data matrix, Co-clustering algorithms often demonstrate surpris- ingly impressive performance improvements over traditional one-sided (row) clust ering techniques. A good clustering of features may be seen as a combinatorial t ransformation of the data matrix, effectively enforcing a form of regularization that may lead to a better clustering of examples (and vice-versa). In many appl ications, partial supervision in the form of a few row labels as well as column labels may be available to potentially assist co-clustering. In this paper, we d evelop two novel semi-supervised multi-class classification algorithms motivated respectively by spectral bipartite graph partitioning and matrix approximation (e.g., non-negative matrix factorization) formulations for co-clustering. These algorithms (i) support dual supervision in the form of labels for both examples and/or features, (ii) provide principled predictive capability on out-of-sample test data, and (iii) arise naturally from the classical Representer theorem appl ied to regularization problems posed on a collection of Reproducing Kernel Hilbe rt Spaces. Empirical results demonstrate the effectiveness and utility of our al gorithms.
****

Tighter Bounds for Structured Estimation
Olivier Chapelle, Chuong B., Choon Teo, Quoc Le, Alex Smola
Large-margin structured estimation methods work by minimizing a convex upper bou nd of loss functions. While they allow for efficient optimization algorithms, th ese convex formulations are not tight and sacrifice the ability to accurately mo del the true loss. We present tighter non-convex bounds based on generalizing th e notion of a ramp loss from binary classification to structured estimation. We show that a small modification of existing optimization algorithms suffices to s olve this modified problem. On structured prediction tasks such as protein seque nce alignment and web page ranking, our algorithm leads to improved accuracy.
****

Multi-Agent Filtering with Infinitely Nested Beliefs
Luke Zettlemoyer, Brian Milch, Leslie Kaelbling
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Beyond Novelty Detection: Incongruent Events, when General and Specific Classifi ers Disagree
Daphna Weinshall, Hynek Hermansky, Alon Zweig, Jie Luo, Holly Jimison, Frank Ohl , Misha Pavel
Unexpected stimuli are a challenge to any machine learning algorithm. Here we id entify distinct types of unexpected events, focusing on 'incongruent events' - w hen 'general level' and 'specific level' classifiers give conflicting prediction s. We define a formal framework for the representation and processing of incongr uent events: starting from the notion of label hierarchy, we show how partial or der on labels can be deduced from such hierarchies. For each event, we compute i ts probability in different ways, based on adjacent levels (according to the par tial order) in the label hierarchy . An incongruent event is an event where the probability computed based on some more specific level (in accordance with the p artial order) is much smaller than the probability computed based on some more g eneral level, leading to conflicting predictions. We derive algorithms to detect incongruent events from different types of hierarchies, corresponding to class membership or part membership. Respectively, we show promising results with real data on two specific problems: Out Of Vocabulary words in speech recognition, a nd the identification of a new sub-class (e.g., the face of a new individual) in audio-visual facial object recognition.

****
Look Ma, No Hands: Analyzing the Monotonic Feature Abstraction for Text Classification
Doug Downey, Oren Etzioni
Is accurate classification possible in the absence of hand-labeled data? This paper introduces the Monotonic Feature (MF) abstraction--where the probability of class membership increases monotonically with the MF's value. The paper proves that when an MF is given, PAC learning is possible with no hand-labeled data under certain assumptions. We argue that MFs arise naturally in a broad range of textual classification applications. On the classic "20 Newsgroups" data set, a learner given an MF and unlabeled data achieves classification accuracy equal to that of a state-of-the-art semi-supervised learner relying on 160 hand-labeled examples. Even when MFs are not given as input, their presence or absence can be determined from a small amount of hand-labeled data, which yields a new semi-supervised learning method that reduces error by 15% on the 20 Newsgroups data.
****
Nonparametric regression and classification with joint sparsity constraints
Han Liu, Larry Wasserman, John Lafferty
We propose new families of models and algorithms for high-dimensional nonparametric learning with joint sparsity constraints. Our approach is based on a regularization method that enforces common sparsity patterns across different function components in a nonparametric additive model. The algorithms employ a coordinate descent approach that is based on a functional soft-thresholding operator. The framework yields several new models, including multi-task sparse additive models, multi-response sparse additive models, and sparse additive multi-category logistic regression. The methods are illustrated with experiments on synthetic data and gene microarray data.
****
Recursive Segmentation and Recognition Templates for 2D Parsing
Leo Zhu, Yuanhao Chen, Yuan Lin, Chenxi Lin, Alan L. Yuille
Language and image understanding are two major goals of arti■cial intelligence which can both be conceptually formulated in terms of parsing the input signal into a hierarchical representation. Natural language researchers have made great progress by exploiting the 1D structure of language to design ef■cient polynomial - time parsing algorithms. By contrast, the two-dimensional nature of images makes it much harder to design ef■cient image parsers and the form of the hierarchical representations is also unclear. Attempts to adapt representations and algorithms from natural language have only been partially successful. In this paper, we propose a Hierarchical Image Model (HIM) for 2D image pars- ing which outputs image segmentation and object recognition. This HIM is rep- resented by recursive segmentation and recognition templates in multiple layers and has advantages for representation, inference, and learning. Firstly, the HIM has a coarse-to-■ne representation which is capable of capturing long-range de- pendency and exploiting different levels of contextual information. Secondly, the structure of the HIM allows us to design a rapid inference algorithm, based on dy- namic programming, which enables us to parse the image rapidly in polynomial time. Thirdly, we can learn the HIM ef■ciently in a discriminative manner from a labeled dataset. We demonstrate that HIM outperforms other state-of-the-art methods by evaluation on the challenging public MSRC image dataset. Finally, we sketch how the HIM architecture can be extended to model more complex image phenomena.
****
Predicting the Geometry of Metal Binding Sites from Protein Sequence
Paolo Frasconi, Andrea Passerini
Metal binding is important for the structural and functional characterization of proteins. Previous prediction efforts have only focused on bonding state, i.e. deciding which protein residues act as metal ligands in some binding site. Identifying the geometry of metal-binding sites, i.e. deciding which residues are jointly involved in the coordination of a metal ion is a new prediction problem that has been never attempted before from protein sequence alone. In this paper, we formulate it in the framework of learning with structured outputs. Our solution

relies on the fact that, from a graph theoretical perspective, metal binding has the algebraic properties of a matroid, enabling the application of greedy algorithms for learning structured outputs. On a data set of 199 non-redundant metalloproteins, we obtained precision/recall levels of 75\%/46\% correct ligand-ion assignments, which improves to 88\%/88\% in the setting where the metal binding state is known.

****

Cell Assemblies in Large Sparse Inhibitory Networks of Biologically Realistic Spiking Neurons

Adam Ponzi, Jeff Wickens

Cell assemblies exhibiting episodes of recurrent coherent activity have been observed in several brain regions including the striatum and hippocampus CA3. Here we address the question of how coherent dynamically switching assemblies appear in large networks of biologically realistic spiking neurons interacting deterministically. We show by numerical simulations of large asymmetric inhibitory networks with fixed external excitatory drive that if the network has intermediate to sparse connectivity, the individual cells are in the vicinity of a bifurcation between a quiescent and firing state and the network inhibition varies slowly on the spiking timescale, then cells form assemblies whose members show strong positive correlation, while members of different assemblies show strong negative correlation. We show that cells and assemblies switch between firing and quiescent states with time durations consistent with a power-law. Our results are in good qualitative agreement with the experimental studies. The deterministic dynamical behaviour is related to winner-less competition shown in small closed loop inhibitory networks with heteroclinic cycles connecting saddle-points.

****

High-dimensional support union recovery in multivariate regression

Guillaume R. Obozinski, Martin J. Wainwright, Michael Jordan

We study the behavior of block (cid:96)1/(cid:96)2 regularization for multivariate regression, where a K-dimensional response vector is regressed upon a ■xed set of p co- variates. The problem of support union recovery is to recover the subset of covariates that are active in at least one of the regression problems. Study- ing this problem under high-dimensional scaling (where the problem parameters as well as sample size n tend to in■nity simultaneously), our main result is to show that exact recovery is possible once the order parameter given by θ(cid:96)1/(cid:96)2(n, p, s) : = n/[2ψ(B∗) log(p − s)] exceeds a critical threshold. Here n is the sample size, p is the ambient dimension of the regression model, s is the size of the union of supports, and ψ(B∗) is a sparsity-overlap function that measures a combination of the sparsities and overlaps of the K-regression coef■cient vectors that constitute the model. This sparsity-overlap function reveals that block (cid:96)1/(cid:96)2 regularization for multivariate regression never harms performance relative to a naive (cid:96)1-approach, and can yield substantial improvements in sample complexity (up to a factor of K) when the regression vectors are suitably orthogonal rela- tive to the design. We complement our theoretical results with simulations that demonstrate the sharpness of the result, even for relatively small problems.

****

Convergence and Rate of Convergence of a Manifold-Based Dimension Reduction Algorithm

Andrew Smith, Hongyuan Zha, Xiao-ming Wu

We study the convergence and the rate of convergence of a local manifold learning algorithm: LTSA [13]. The main technical tool is the perturbation analysis on the linear invariant subspace that corresponds to the solution of LTSA. We derive a worst-case upper bound of errors for LTSA which naturally leads to a convergence result. We then derive the rate of convergence for LTSA in a special case.

****

A Convex Upper Bound on the Log-Partition Function for Binary Distributions

Laurent Ghaoui, Assane Gueye

We consider the problem of bounding from above the log-partition function corresponding to second-order Ising models for binary distributions. We introduce a ne

w bound, the cardinality bound, which can be computed via convex optimization. T
he corresponding error on the logpartition function is bounded above by twice th
e distance, in model parameter space, to a class of "standard" Ising models, for
 which variable inter-dependence is described via a simple mean field term. In t
he context of maximum-likelihood, using the new bound instead of the exact log-p
artition function, while constraining the distance to the class of standard Isin
g models, leads not only to a good approximation to the log-partition function,
but also to a model that is parsimonious, and easily interpretable. We compare o
ur bound with the log-determinant bound introduced by Wainwright and Jordan (200
6), and show that when the l1 -norm of the model parameter vector is small enoug
h, the latter is outperformed by the new bound.
****

Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Model
s

Tong Zhang

Consider linear prediction models where the target function is a sparse linear c
ombination of a set of basis functions. We are interested in the problem of iden
tifying those basis functions with non-zero coefficients and reconstructing the
target function from noisy observations. Two heuristics that are widely used in
practice are forward and backward greedy algorithms. First, we show that neither
 idea is adequate. Second, we propose a novel combination that is based on the f
orward greedy algorithm but takes backward steps adaptively whenever beneficial.
 We prove strong theoretical results showing that this procedure is effective in
 learning sparse representations. Experimental results support our theory.
****

Reconciling Real Scores with Binary Comparisons: A New Logistic Based Model for
Ranking

Nir Ailon

The problem of ranking arises ubiquitously in almost every aspect of life, and i
n particular in Machine Learning/Information Retrieval. A statistical model for
ranking predicts how humans rank subsets V of some universe U . In this work we
define a statistical model for ranking that satisfies certain desirable properti
es. The model automatically gives rise to a logistic regression based approach t
o learning how to rank, for which the score and comparison based approaches are
dual views. This offers a new generative approach to ranking which can be used f
or IR. There are two main contexts for this work. The first is the theory of eco
nometrics and study of statistical models explaining human choice of alternative
s. In this context, we will compare our model with other well known models. The
second context is the problem of ranking in machine learning, usually arising in
 the context of information retrieval. Here, much work has been done in the disc
riminative setting, where different heuristics are used to define ranking risk f
unctions. Our model is built rigorously and axiomatically based on very simple d
esirable properties defined locally for comparisons, and automatically implies t
he existence of a global score function serving as a natural model parameter whi
ch can be efficiently fitted to pairwise comparison judgment data by solving a c
onvex optimization problem.
****

Theory of matching pursuit

Zakria Hussain, John Shawe-taylor

We analyse matching pursuit for kernel principal components analysis by proving
that the sparse subspace it produces is a sample compression scheme. We show tha
t this bound is tighter than the KPCA bound of Shawe-Taylor et al swck-05 and hi
ghly predictive of the size of the subspace needed to capture most of the varian
ce in the data. We analyse a second matching pursuit algorithm called kernel mat
ching pursuit (KMP) which does not correspond to a sample compression scheme. Ho
wever, we give a novel bound that views the choice of subspace of the KMP algori
thm as a compression scheme and hence provide a VC bound to upper bound its futu
re loss. Finally we describe how the same bound can be applied to other matching
 pursuit related algorithms.
****

Policy Search for Motor Primitives in Robotics
Jens Kober, Jan Peters
Many motor skills in humanoid robotics can be learned using parametrized motor p
rimitives as done in imitation learning. However, most interesting motor learnin
g problems are high-dimensional reinforcement learning problems often beyond the
 reach of current methods. In this paper, we extend previous work on policy lear
ning from the immediate reward case to episodic reinforcement learning. We show
 that this results into a general, common framework also connected to policy grad
ient methods and yielding a novel algorithm for policy learning by assuming a fo
rm of exploration that is particularly well-suited for dynamic motor primitives.
 The resulting algorithm is an EM-inspired algorithm applicable in complex motor
 learning tasks. We compare this algorithm to alternative parametrized policy se
arch methods and show that it outperforms previous methods. We apply it in the c
ontext of motor learning and show that it can learn a complex Ball-in-a-Cup task
 using a real Barrett WAM robot arm.
****
Mortal Multi-Armed Bandits
Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, Eli Upfal
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.
****
Adapting to a Market Shock: Optimal Sequential Market-Making
Sanmay Das, Malik Magdon-Ismail
We study the profit-maximization problem of a monopolistic market-maker who sets
 two-sided prices in an asset market. The sequential decision problem is hard to
 solve because the state space is a function. We demonstrate that the belief sta
te is well approximated by a Gaussian distribution. We prove a key monotonicity
property of the Gaussian state update which makes the problem tractable, yieldin
g the first optimal sequential market-making algorithm in an established model.
The algorithm leads to a surprising insight: an optimal monopolist can provide m
ore liquidity than perfectly competitive market-makers in periods of extreme unc
ertainty, because a monopolist is willing to absorb initial losses in order to l
earn a new valuation rapidly so she can extract higher profits later.
****
DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification
Simon Lacoste-Julien, Fei Sha, Michael Jordan
Probabilistic topic models (and their extensions) have become popular as models
of latent structures in collections of text documents or images. These models ar
e usually treated as generative models and trained using maximum likelihood esti
mation, an approach which may be suboptimal in the context of an overall classif
ication problem. In this paper, we describe DiscLDA, a discriminative learning f
ramework for such models as Latent Dirichlet Allocation (LDA) in the setting of
dimensionality reduction with supervised side information. In DiscLDA, a class-d
ependent linear transformation is introduced on the topic mixture proportions. T
his parameter is estimated by maximizing the conditional likelihood using Monte
Carlo EM. By using the transformed topic mixture proportions as a new representa
tion of documents, we obtain a supervised dimensionality reduction algorithm tha
t uncovers the latent structure in a document collection while preserving predic
tive power for the task of classification. We compare the predictive power of th
e latent structure of DiscLDA with unsupervised LDA on the 20 Newsgroup ocument
classification task.
****
Understanding Brain Connectivity Patterns during Motor Imagery for Brain-Compute
r Interfacing
Moritz Grosse-wentrup
EEG connectivity measures could provide a new type of feature space for inferrin
g a subject's intention in Brain-Computer Interfaces (BCIs). However, very littl
e is known on EEG connectivity patterns for BCIs. In this study, EEG connectivit

y during motor imagery (MI) of the left and right is investigated in a broad fre
quency range across the whole scalp by combining Beamforming with Transfer Entro
py and taking into account possible volume conduction effects. Observed connecti
vity patterns indicate that modulation intentionally induced by MI is strongest
in the gamma-band, i.e., above 35 Hz. Furthermore, modulation between MI and res
t is found to be more pronounced than between MI of different hands. This is in
contrast to results on MI obtained with bandpower features, and might provide an
 explanation for the so far only moderate success of connectivity features in BC
Is. It is concluded that future studies on connectivity based BCIs should focus
on high frequency bands and consider experimental paradigms that maximally vary
cognitive demands between conditions.
****

Rademacher Complexity Bounds for Non-I.I.D. Processes
Mehryar Mohri, Afshin Rostamizadeh
This paper presents the first data-dependent generalization bounds for non-i.i.d
. settings based on the notion of Rademacher complexity. Our bounds extend to th
e non-i.i.d. case existing Rademacher complexity bounds derived for the i.i.d. s
etting. These bounds provide a strict generalization of the ones found in the i.
i.d. case, and can also be used within the standard i.i.d. scenario. They apply
to the standard scenario of beta-mixing stationary sequences examined in many pr
evious studies of non-i.i.d. settings and benefit form the crucial advantages of
 Rademacher complexity over other measures of the complexity of hypothesis class
es. In particular, they are data-dependent and measure the complexity of a class
 of hypotheses based on the training sample. The empirical Rademacher complexity
 can be estimated from finite samples and lead to tighter bounds.
****

A rational model of preference learning and choice prediction by children
Christopher Lucas, Thomas Griffiths, Fei Xu, Christine Fawcett
Young children demonstrate the ability to make inferences about the preferences
of other agents based on their choices. However, there exists no overarching acc
ount of what children are doing when they learn about preferences or how they us
e that knowledge. We use a rational model of preference learning, drawing on ide
as from economics and computer science, to explain the behavior of children in s
everal recent experiments. Specifically, we show how a simple econometric model
can be extended to capture two- to four-year-oldsâ■■ use of statistical informat
ion in inferring preferences, and their generalization of these preferences.
****

An ideal observer model of infant object perception
Charles Kemp, Fei Xu
Before the age of 4 months, infants make inductive inferences about the motions
of physical objects. Developmental psychologists have provided verbal accounts o
f the knowledge that supports these inferences, but often these accounts focus o
n categorical rather than probabilistic principles. We propose that infant objec
t perception is guided in part by probabilistic principles like persistence: thi
ngs tend to remain the same, and when they change they do so gradually. To illus
trate this idea, we develop an ideal observer model that includes probabilistic
formulations of rigidity and inertia. Like previous researchers, we suggest that
 rigid motions are expected from an early age, but we challenge the previous cla
im that expectations consistent with inertia are relatively slow to develop (Spe
lke et al., 1992). We support these arguments by modeling four experiments from
the developmental literature.
****

Empirical performance maximization for linear rank statistics
Stéphan Clémençcon, Nicolas Vayatis
The ROC curve is known to be the golden standard for measuring performance of a
test/scoring statistic regarding its capacity of discrimination between two popu
lations in a wide variety of applications, ranging from anomaly detection in sig
nal processing to information retrieval, through medical diagnosis. Most practic
al performance measures used in scoring applications such as the AUC, the local
AUC, the p-norm push, the DCG and others, can be seen as summaries of the ROC cu

rve. This paper highlights the fact that many of these empirical criteria can be expressed as (conditional) linear rank statistics. We investigate the properties of empirical maximizers of such performance criteria and provide preliminary results for the concentration properties of a novel class of random variables that we will call a linear rank process.
****

Resolution Limits of Sparse Coding in High Dimensions
Sundeep Rangan, Vivek Goyal, Alyson K. Fletcher
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.
****

Privacy-preserving logistic regression
Kamalika Chaudhuri, Claire Monteleoni
This paper addresses the important tradeoff between privacy and learnability, when designing algorithms for learning from private databases. First we apply an idea of Dwork et al. to design a specific privacy-preserving machine learning algorithm, logistic regression. This involves bounding the sensitivity of logistic regression, and perturbing the learned classifier with noise proportional to the sensitivity. Noting that the approach of Dwork et al. has limitations when applied to other machine learning algorithms, we then present another privacy-preserving logistic regression algorithm. The algorithm is based on solving a perturbed objective, and does not depend on the sensitivity. We prove that our algorithm preserves privacy in the model due to Dwork et al., and we provide a learning performance guarantee. Our work also reveals an interesting connection between regularization and privacy.
****

Efficient Exact Inference in Planar Ising Models
Nicol Schraudolph, Dmitry Kamenetsky
We present polynomial-time algorithms for the exact computation of lowest- energy states, worst margin violators, partition functions, and marginals in binary undirected graphical models. Our approach provides an interesting alternative to the well-known graph cut paradigm in that it does not impose any submodularity constraints; instead we require planarity to establish a correspondence with perfect matchings in an expanded dual graph. Maximum-margin parameter estimation for a boundary detection task shows our approach to be ef■cient and effective.
****

Deflation Methods for Sparse PCA
Lester Mackey
In analogy to the PCA setting, the sparse PCA problem is often solved by iteratively alternating between two subtasks: cardinality-constrained rank-one variance maximization and matrix deflation. While the former has received a great deal of attention in the literature, the latter is seldom analyzed and is typically borrowed without justification from the PCA context. In this work, we demonstrate that the standard PCA deflation procedure is seldom appropriate for the sparse PCA setting. To rectify the situation, we first develop several heuristic deflation alternatives with more desirable properties. We then reformulate the sparse PCA optimization problem to explicitly reflect the maximum additional variance objective on each round. The result is a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets.
****

Mixed Membership Stochastic Blockmodels
Edo M. Airoldi, David Blei, Stephen Fienberg, Eric Xing
Observations consisting of measurements on relationships for pairs of objects arise in many settings, such as protein interaction and gene regulatory networks, collections of author-recipient email, and social networks. Analyzing such data with probabilisic models can be delicate because the simple exchangeability assumptions underlying many boilerplate models no longer hold. In this paper, we describe a class of latent variable models of such data called Mixed Membership Sto

chastic Blockmodels. This model extends blockmodels for relational data to ones which capture mixed membership latent relational structure, thus providing an object-specific low-dimensional representation. We develop a general variational inference algorithm for fast approximate posterior inference. We explore applications to social networks and protein interaction networks.
****

Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes
Erik Sudderth, Michael Jordan
We develop a statistical framework for the simultaneous, unsupervised segmentation and discovery of visual object categories from image databases. Examining a large set of manually segmented scenes, we use chi--square tests to show that object frequencies and segment sizes both follow power law distributions, which are well modeled by the Pitman--Yor (PY) process. This nonparametric prior distribution leads to learning algorithms which discover an unknown set of objects, and segmentation methods which automatically adapt their resolution to each image. Generalizing previous applications of PY processes, we use Gaussian processes to discover spatially contiguous segments which respect image boundaries. Using a novel family of variational approximations, our approach produces segmentations which compare favorably to state--of--the--art methods, while simultaneously discovering categories shared among natural scenes.
****

Shape-Based Object Localization for Descriptive Classification
Geremy Heitz, Gal Elidan, Benjamin Packer, Daphne Koller
Discriminative tasks, including object categorization and detection, are central components of high-level computer vision. Sometimes, however, we are interested in more refined aspects of the object in an image, such as pose or particular regions. In this paper we develop a method (LOOPS) for learning a shape and image feature model that can be trained on a particular object class, and used to outline instances of the class in novel images. Furthermore, while the training data consists of uncorresponded outlines, the resulting LOOPS model contains a set of landmark points that appear consistently across instances, and can be accurately localized in an image. Our model achieves state-of-the-art results in precisely outlining objects that exhibit large deformations and articulations in cluttered natural images. These localizations can then be used to address a range of tasks, including descriptive classification, search, and clustering.
****

Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform
Guangzhi Cao, Charles Bouman
Covariance estimation for high dimensional vectors is a classically difficult problem in statistical analysis and machine learning due to limited sample size. In this paper, we propose a new approach to covariance estimation, which is based on constrained maximum likelihood (ML) estimation of the covariance. Specifically, the covariance is constrained to have an eigen decomposition which can be represented as a sparse matrix transform (SMT). The SMT is formed by a product of pairwise coordinate rotations known as Givens rotations. Using this framework, the covariance can be efficiently estimated using greedy minimization of the log likelihood function, and the number of Givens rotations can be efficiently computed using a cross-validation procedure. The estimator obtained using this method is always positive definite and well-conditioned even with limited sample size. Experiments on hyperspectral data show that SMT covariance estimation results in consistently better estimates of the covariance for a variety of different classes and sample sizes compared to traditional shrinkage estimators.
****

Characterizing neural dependencies with copula models
Pietro Berkes, Frank Wood, Jonathan Pillow
The coding of information by neural populations depends critically on the statistical dependencies between neuronal responses. However, there is no simple model that combines the observations that (1) marginal distributions over single-neuron spike counts are often approximately Poisson; and (2) joint distributions ove

r the responses of multiple neurons are often strongly dependent. Here, we show that both marginal and joint properties of neural responses can be captured using Poisson copula models. Copulas are joint distributions that allow random variables with arbitrary marginals to be combined while incorporating arbitrary dependencies between them. Different copulas capture different kinds of dependencies, allowing for a richer and more detailed description of dependencies than traditional summary statistics, such as correlation coefficients. We explore a variety of Poisson copula models for joint neural response distributions, and derive an efficient maximum likelihood procedure for estimating them. We apply these models to neuronal data collected in and macaque motor cortex, and quantify the improvement in coding accuracy afforded by incorporating the dependency structure between pairs of neurons.
****

Learning with Consistency between Inductive Functions and Kernels
Haixuan Yang, Irwin King, Michael Lyu
Regularized Least Squares (RLS) algorithms have the ability to avoid over-fitting problems and to express solutions as kernel expansions. However, we observe that the current RLS algorithms cannot provide a satisfactory interpretation even on a constant function. On the other hand, while kernel-based algorithms have been developed in such a tendency that almost all learning algorithms are kernelized or being kernelized, a basic fact is often ignored: The learned function from the data and the kernel fits the data well, but may not be consistent with the kernel. Based on these considerations and on the intuition that a good kernel-based inductive function should be consistent with both the data and the kernel, a novel learning scheme is proposed. The advantages of this scheme lie in its corresponding Representer Theorem, its strong interpretation ability about what kind of functions should not be penalized, and its promising accuracy improvements shown in a number of experiments. Furthermore, we provide a detailed technical description about heat kernels, which serves as an example for the readers to apply similar techniques for other kernels. Our work provides a preliminary step in a new direction to explore the varying consistency between inductive functions and kernels under various distributions.
****

Syntactic Topic Models
Jordan Boyd-graber, David Blei
We develop \name\ (STM), a nonparametric Bayesian model of parsed documents. \Shortname\ generates words that are both thematically and syntactically constrained, which combines the semantic insights of topic models with the syntactic information available from parse trees. Each word of a sentence is generated by a distribution that combines document-specific topic weights and parse-tree specific syntactic transitions. Words are assumed generated in an order that respects the parse tree. We derive an approximate posterior inference method based on variational methods for hierarchical Dirichlet processes, and we report qualitative and quantitative results on both synthetic data and hand-parsed documents.
****

A general framework for investigating how far the decoding process in the brain can be simplified
Masafumi Oizumi, Toshiyuki Ishii, Kazuya Ishibashi, Toshihiko Hosoya, Masato Okada
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.
****

Signal-to-Noise Ratio Analysis of Policy Gradient Algorithms
John Roberts, Russ Tedrake
Policy gradient (PG) reinforcement learning algorithms have strong (local) convergence guarantees, but their learning performance is typically limited by a large variance in the estimate of the gradient. In this paper, we formulate the variance reduction problem by describing a signal-to-noise ratio (SNR) for policy gr

adient algorithms, and evaluate this SNR carefully for the popular Weight Pertur bation (WP) algorithm. We confirm that SNR is a good predictor of long-term lear ning performance, and that in our episodic formulation, the cost-to-go function is indeed the optimal baseline. We then propose two modifications to traditional model-free policy gradient algorithms in order to optimize the SNR. First, we e xamine WP using anisotropic sampling distributions, which introduces a bias into the update but increases the SNR; this bias can be interpretted as following th e natural gradient of the cost function. Second, we show that non-Gaussian distr ibutions can also increase the SNR, and argue that the optimal isotropic distrib ution is a â■■shellâ■■ distribution with a constant magnitude and uniform distri bution in direction. We demonstrate that both modifications produce substantial improvements in learning performance in challenging policy gradient experiments.
****

Artificial Olfactory Brain for Mixture Identification

Mehmet Muezzinoglu, Alexander Vergara, Ramon Huerta, Thomas Nowotny, Nikolai Rul kov, Henry Abarbanel, Allen Selverston, Mikhail Rabinovich

The odor transduction process has a large time constant and is susceptible to va rious types of noise. Therefore, the olfactory code at the sensor/receptor level is in general a slow and highly variable indicator of the input odor in both na tural and artificial situations. Insects overcome this problem by using a neuron al device in their Antennal Lobe (AL), which transforms the identity code of olf actory receptors to a spatio-temporal code. This transformation improves the dec ision of the Mushroom Bodies (MBs), the subsequent classifier, in both speed and accuracy.Here we propose a rate model based on two intrinsic mechanisms in the insect AL, namely integration and inhibition. Then we present a MB classifier mo del that resembles the sparse and random structure of insect MB. A local Hebbian learning procedure governs the plasticity in the model. These formulations not only help to understand the signal conditioning and classification methods of in sect olfactory systems, but also can be leveraged in synthetic problems. Among t hem, we consider here the discrimination of odor mixtures from pure odors. We sh ow on a set of records from metal-oxide gas sensors that the cascade of these tw o new models facilitates fast and accurate discrimination of even highly imbalan ced mixtures from pure odors.
****

Robust Kernel Principal Component Analysis

Minh Nguyen, Fernando Torre

Kernel Principal Component Analysis (KPCA) is a popular generalization of linear PCA that allows non-linear feature extraction. In KPCA, data in the input space is mapped to higher (usually) dimensional feature space where the data can be l inearly modeled. The feature space is typically induced implicitly by a kernel f unction, and linear PCA in the feature space is performed via the kernel trick. However, due to the implicitness of the feature space, some extensions of PCA su ch as robust PCA cannot be directly generalized to KPCA. This paper presents a t echnique to overcome this problem, and extends it to a unified framework for tre ating noise, missing data, and outliers in KPCA. Our method is based on a novel cost function to perform inference in KPCA. Extensive experiments, in both synth etic and real data, show that our algorithm outperforms existing methods.
****

Relative Margin Machines

Tony Jebara, Pannagadatta Shivaswamy

In classification problems, Support Vector Machines maximize the margin of separ ation between two classes. While the paradigm has been successful, the solution obtained by SVMs is dominated by the directions with large data spread and biase d to separate the classes by cutting along large spread directions. This article proposes a novel formulation to overcome such sensitivity and maximizes the mar gin relative to the spread of the data. The proposed formulation can be efficien tly solved and experiments on digit datasets show drastic performance improvemen ts over SVMs.
****

Bayesian Kernel Shaping for Learning Control

Jo-anne Ting, Mrinal Kalakrishnan, Sethu Vijayakumar, Stefan Schaal
In kernel-based regression learning, optimizing each kernel individually is useful when the data density, curvature of regression surfaces (or decision boundaries) or magnitude of output noise (i.e., heteroscedasticity) varies spatially. Unfortunately, it presents a complex computational problem as the danger of overfitting is high and the individual optimization of every kernel in a learning system may be overly expensive due to the introduction of too many open learning parameters. Previous work has suggested gradient descent techniques or complex statistical hypothesis methods for local kernel shaping, typically requiring some amount of manual tuning of meta parameters. In this paper, we focus on nonparametric regression and introduce a Bayesian formulation that, with the help of variational approximations, results in an EM-like algorithm for simultaneous estimation of regression and kernel parameters. The algorithm is computationally efficient (suitable for large data sets), requires no sampling, automatically rejects outliers and has only one prior to be specified. It can be used for nonparametric regression with local polynomials or as a novel method to achieve nonstationary regression with Gaussian Processes. Our methods are particularly useful for learning control, where reliable estimation of local tangent planes is essential for adaptive controllers and reinforcement learning. We evaluate our methods on several synthetic data sets and on an actual robot which learns a task-level control law.
****

An Extended Level Method for Efficient Multiple Kernel Learning
Zenglin Xu, Rong Jin, Irwin King, Michael Lyu
We consider the problem of multiple kernel learning (MKL), which can be formulated as a convex-concave problem. In the past, two efficient methods, i.e., Semi-Infinite Linear Programming (SILP) and Subgradient Descent (SD), have been proposed for large-scale multiple kernel learning. Despite their success, both methods have their own shortcomings: (a) the SD method utilizes the gradient of only the current solution, and (b) the SILP method does not regularize the approximate solution obtained from the cutting plane model. In this work, we extend the level method, which was originally designed for optimizing non-smooth objective functions, to convex-concave optimization, and apply it to multiple kernel learning. The extended level method overcomes the drawbacks of SILP and SD by exploiting all the gradients computed in past iterations and by regularizing the solution via a projection to a level set. Empirical study with eight UCI datasets shows that the extended level method can significantly improve efficiency by saving on average 91.9% of computational time over the SILP method and 70.3% over the SD method.
****

Sparse Online Learning via Truncated Gradient
John Langford, Lihong Li, Tong Zhang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Bayesian Model of Behaviour in Economic Games
Debajyoti Ray, Brooks King-casas, P. Montague, Peter Dayan
Classical Game Theoretic approaches that make strong rationality assumptions have difficulty modeling observed behaviour in Economic games of human subjects. We investigate the role of finite levels of iterated reasoning and non-selfish utility functions in a Partially Observable Markov Decision Process model that incorporates Game Theoretic notions of interactivity. Our generative model captures a broad class of characteristic behaviours in a multi-round Investment game. We invert the generative process for a recognition model that is used to classify 200 subjects playing an Investor-Trustee game against randomly matched opponents.
****

Skill Characterization Based on Betweenness
Özgür ■im■ek, Andrew Barto

We present a characterization of a useful class of skills based on a graphical r
epresentation of an agent's interaction with its environment. Our characterizati
on uses betweenness, a measure of centrality on graphs. It may be used directly
to form a set of skills suitable for a given environment. More importantly, it s
erves as a useful guide for developing online, incremental skill discovery algor
ithms that do not rely on knowing or representing the environment graph in its e
ntirety.
****

Improved Moves for Truncated Convex Models
Philip Torr, M. Kumar
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.
****

Kernelized Sorting
Novi Quadrianto, Le Song, Alex Smola
Object matching is a fundamental operation in data analysis. It typically requir
es the definition of a similarity measure between the classes of objects to be m
atched. Instead, we develop an approach which is able to perform matching by req
uiring a similarity measure only within each of the classes. This is achieved by
 maximizing the dependency between matched pairs of observations by means of the
 Hilbert Schmidt Independence Criterion. This problem can be cast as one of maxi
mizing a quadratic assignment problem with special structure and we present a si
mple algorithm for finding a locally optimal solution.
****

Posterior Consistency of the Silverman g-prior in Bayesian Model Choice
Zhihua Zhang, Michael Jordan, Dit-Yan Yeung
Kernel supervised learning methods can be unified by utilizing the tools from re
gularization theory. The duality between regularization and prior leads to inter
preting regularization methods in terms of maximum a posteriori estimation and h
as motivated Bayesian interpretations of kernel methods. In this paper we pursue
 a Bayesian interpretation of sparsity in the kernel setting by making use of a
mixture of a point-mass distribution and prior that we refer to as ``Silverman's
 g-prior.'' We provide a theoretical analysis of the posterior consistency of a
Bayesian model choice procedure based on this prior. We also establish the asymp
totic relationship between this procedure and the Bayesian information criterion
.
****

Nonparametric Bayesian Learning of Switching Linear Dynamical Systems
Emily Fox, Erik Sudderth, Michael Jordan, Alan Willsky
Many nonlinear dynamical phenomena can be effectively modeled by a system that s
witches among a set of conditionally linear dynamical modes. We consider two suc
h models: the switching linear dynamical system (SLDS) and the switching vector
autoregressive (VAR) process. In this paper, we present a nonparametric approach
 to the learning of an unknown number of persistent, smooth dynamical modes by u
tilizing a hierarchical Dirichlet process prior. We develop a sampling algorithm
 that combines a truncated approximation to the Dirichlet process with an effici
ent joint sampling of the mode and state sequences. The utility and flexibility
of our model are demonstrated on synthetic data, sequences of dancing honey bees
, and the IBOVESPA stock index.
****

Learning to Use Working Memory in Partially Observable Environments through Dopa
minergic Reinforcement
Michael Todd, Yael Niv, Jonathan D. Cohen
Working memory is a central topic of cognitive neuroscience because it is critic
al for solving real world problems in which information from multiple temporally
 distant sources must be combined to generate appropriate behavior. However, an
often neglected fact is that learning to use working memory effectively is itsel
f a difficult problem. The Gating" framework is a collection of psychological mo

dels that show how dopamine can train the basal ganglia and prefrontal cortex to form useful working memory representations in certain types of problems. We bring together gating with ideas from machine learning about using finite memory systems in more general problems. Thus we present a normative Gating model that learns, by online temporal difference methods, to use working memory to maximize discounted future rewards in general partially observable settings. The model successfully solves a benchmark working memory problem, and exhibits limitations similar to those observed in human experiments. Moreover, the model introduces a concise, normative definition of high level cognitive concepts such as working memory and cognitive control in terms of maximizing discounted future rewards."
****

Optimization on a Budget: A Reinforcement Learning Approach
Paul Ruvolo, Ian Fasel, Javier Movellan
Many popular optimization algorithms, like the Levenberg-Marquardt algorithm (LMA), use heuristic-based controllers'' that modulate the behavior of the optimizer during the optimization process. For example, in the LMA a damping parameter is dynamically modified based on a set rules that were developed using various heuristic arguments. Reinforcement learning (RL) is a machine learning approach to learn optimal controllers by examples and thus is an obvious candidate to improve the heuristic-based controllers implicit in the most popular and heavily used optimization algorithms. Improving the performance of off-the-shelf optimizers is particularly important for time-constrained optimization problems. For example the LMA algorithm has become popular for many real-time computer vision problems, including object tracking from video, where only a small amount of time can be allocated to the optimizer on each incoming video frame. Here we show that a popular modern reinforcement learning technique using a very simply state space can dramatically improve the performance of general purpose optimizers, like the LMA. Most surprisingly the controllers learned for a particular domain appear to work very well also on very different optimization domains. For example we used RL methods to train a new controller for the damping parameter of the LMA. This controller was trained on a collection of classic, relatively small, non-linear regression problems. The modified LMA performed better than the standard LMA on these problems. Most surprisingly, it also dramatically outperformed the standard LMA on a difficult large scale computer vision problem for which it had not been trained before. Thus the controller appeared to have extracted control rules that were not just domain specific but generalized across a wide range of optimization domains."
****

Dependent Dirichlet Process Spike Sorting
Jan Gasthaus, Frank Wood, Dilan Gorur, Yee Teh
In this paper we propose a new incremental spike sorting model that automatically eliminates refractory period violations, accounts for action potential waveform drift, and can handle appearance" and "disappearance" of neurons. Our approach is to augment a known time-varying Dirichlet process that ties together a sequence of infinite Gaussian mixture models, one per action potential waveform observation, with an interspike-interval-dependent likelihood that prohibits refractory period violations. We demonstrate this model by showing results from sorting two publicly available neural data recordings for which the a partial ground truth labeling is known."
****

The Recurrent Temporal Restricted Boltzmann Machine
Ilya Sutskever, Geoffrey E. Hinton, Graham W. Taylor
The Temporal Restricted Boltzmann Machine (TRBM) is a probabilistic model for sequences that is able to successfully model (i.e., generate nice-looking samples of) several very high dimensional sequences, such as motion capture data and the pixels of low resolution videos of balls bouncing in a box. The major disadvantage of the TRBM is that exact inference is extremely hard, since even computing a Gibbs update for a single variable of the posterior is exponentially expensive. This difficulty has necessitated the use of a heuristic inference procedure, that nonetheless was accurate enough for successful learning. In this paper we in

troduce the Recurrent TRBM, which is a very slight modification of the TRBM for which exact inference is very easy and exact gradient learning is almost tractable. We demonstrate that the RTRBM is better than an analogous TRBM at generating motion capture and videos of bouncing balls.
****

Short-Term Depression in VLSI Stochastic Synapse
Peng Xu, Timothy Horiuchi, Pamela Abshire
We report a compact realization of short-term depression (STD) in a VLSI stochastic synapse. The behavior of the circuit is based on a subtractive single release model of STD. Experimental results agree well with simulation and exhibit expected STD behavior: the transmitted spike train has negative autocorrelation and lower power spectral density at low frequencies which can remove redundancy in the input spike train, and the mean transmission probability is inversely proportional to the input spike rate which has been suggested as an automatic gain control mechanism in neural systems. The dynamic stochastic synapse could potentially be a powerful addition to existing deterministic VLSI spiking neural systems.
****

Generative and Discriminative Learning with Unknown Labeling Bias
Steven Phillips, Miroslav Dudík
We apply robust Bayesian decision theory to improve both generative and discriminative learners under bias in class proportions in labeled training data, when the true class proportions are unknown. For the generative case, we derive an entropy-based weighting that maximizes expected log likelihood under the worst-case true class proportions. For the discriminative case, we derive a multinomial logistic model that minimizes worst-case conditional log loss. We apply our theory to the modeling of species geographic distributions from presence data, an extreme case of label bias since there is no absence data. On a benchmark dataset, we find that entropy-based weighting offers an improvement over constant estimates of class proportions, consistently reducing log loss on unbiased test data.
****

Large Margin Taxonomy Embedding for Document Categorization
Kilian Q. Weinberger, Olivier Chapelle
Applications of multi-class classification, such as document categorization, often appear in cost-sensitive settings. Recent work has significantly improved the state of the art by moving beyond ``flat'' classification through incorporation of class hierarchies [Cai and Hoffman 04]. We present a novel algorithm that goes beyond hierarchical classification and estimates the latent semantic space that underlies the class hierarchy. In this space, each class is represented by a prototype and classification is done with the simple nearest neighbor rule. The optimization of the semantic space incorporates large margin constraints that ensure that for each instance the correct class prototype is closer than any other. We show that our optimization is convex and can be solved efficiently for large data sets. Experiments on the OHSUMED medical journal data base yield state-of-the-art results on topic categorization.
****

Modeling the effects of memory on human online sentence processing with particle filters
Roger Levy, Florencia Reali, Thomas Griffiths
Language comprehension in humans is significantly constrained by memory, yet rapid, highly incremental, and capable of utilizing a wide range of contextual information to resolve ambiguity and form expectations about future input. In contrast, most of the leading psycholinguistic models and fielded algorithms for natural language parsing are non-incremental, have run time superlinear in input length, and/or enforce structural locality constraints on probabilistic dependencies between events. We present a new limited-memory model of sentence comprehension which involves an adaptation of the particle filter, a sequential Monte Carlo method, to the problem of incremental parsing. We show that this model can reproduce classic results in online sentence comprehension, and that it naturally provides the first rational account of an outstanding problem in psycholinguistics, in which the preferred alternative in a syntactic ambiguity seems to grow more a

ttractive over time even in the absence of strong disambiguating information.
****
Clusters and Coarse Partitions in LP Relaxations
David Sontag, Amir Globerson, Tommi Jaakkola
We propose a new class of consistency constraints for Linear Programming (LP) re laxations for finding the most probable (MAP) configuration in graphical models. Usual cluster-based LP relaxations enforce joint consistency of the beliefs of a cluster of variables, with computational cost increasing exponentially with th e size of the clusters. By partitioning the state space of a cluster and enforci ng consistency only across partitions, we obtain a class of constraints which, a lthough less tight, are computationally feasible for large clusters. We show how to solve the cluster selection and partitioning problem monotonically in the du al LP, using the current beliefs to guide these choices. We obtain a dual messag e-passing algorithm and apply it to protein design problems where the variables have large state spaces and the usual cluster-based relaxations are very costly.
****
Bounds on marginal probability distributions
Joris M. Mooij, Hilbert Kappen
We propose a novel bound on single-variable marginal probability distributions i n factor graphs with discrete variables. The bound is obtained by propagating bo unds (convex sets of probability distributions) over a subtree of the factor gra ph, rooted in the variable of interest. By construction, the method not only bou nds the exact marginal probability distribution of a variable, but also its appr oximate Belief Propagation marginal (``belief''). Thus, apart from providing a p ractical means to calculate bounds on marginals, our contribution also lies in p roviding a better understanding of the error made by Belief Propagation. We show that our bound outperforms the state-of-the-art on some inference problems aris ing in medical diagnosis.
****
Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Al location
Indraneel Mukherjee, David Blei
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.
****
Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning
Francis Bach
For supervised and unsupervised learning, positive definite kernels allow to use large and potentially infinite dimensional feature spaces with a computational cost that only depends on the number of observations. This is usually done throu gh the penalization of predictor functions by Euclidean or Hilbertian norms. In this paper, we explore penalizing by sparsity-inducing norms such as the L1-norm or the block L1-norm. We assume that the kernel decomposes into a large sum of individual basis kernels which can be embedded in a directed acyclic graph; we s how that it is then possible to perform kernel selection through a hierarchical multiple kernel learning framework, in polynomial time in the number of selected kernels. This framework is naturally applied to non linear variable selection; our extensive simulations on synthetic datasets and datasets from the UCI reposi tory show that efficiently exploring the large feature space through sparsity-in ducing norms leads to state-of-the-art predictive performance.
****
MCBoost: Multiple Classifier Boosting for Perceptual Co-clustering of Images and Visual Features
Tae-kyun Kim, Roberto Cipolla
We present a new co-clustering problem of images and visual features. The proble m involves a set of non-object images in addition to a set of object images and features to be co-clustered. Co-clustering is performed in a way of maximising d iscrimination of object images from non-object images, thus emphasizing discrimi

native features. This provides a way of obtaining perceptual joint-clusters of o bject images and features. We tackle the problem by simultaneously boosting mult iple strong classifiers which compete for images by their expertise. Each boosti ng classifier is an aggregation of weak-learners, i.e. simple visual features. T he obtained classifiers are useful for multi-category and multi-view object dete ction tasks. Experiments on a set of pedestrian images and a face data set demon strate that the method yields intuitive image clusters with associated features and is much superior to conventional boosting classifiers in object detection ta sks.

****

Load and Attentional Bayes

Peter Dayan

Selective attention is a most intensively studied psychological phenomenon, rife with theoretical suggestions and schisms. A critical idea is that of limited ca pacity, the allocation of which has produced half a century's worth of conflict about such phenomena as early and late selection. An influential resolution of t his debate is based on the notion of perceptual load (Lavie, 2005, TICS, 9: 75), which suggests that low-load, easy tasks, because they underuse the total capac ity of attention, mandatorily lead to the processing of stimuli that are irrelev ant to the current attentional set; whereas high-load, difficult tasks grab all resources for themselves, leaving distractors high and dry. We argue that this t heory presents a challenge to Bayesian theories of attention, and suggest an alt ernative, statistical, account of key supporting data.

****

Dependence of Orientation Tuning on Recurrent Excitation and Inhibition in a Net work Model of V1

Klaus Wimmer, Marcel Stimberg, Robert Martin, Lars Schwabe, Jorge Mariño, James Schummers, David Lyon, Mriganka Sur, Klaus Obermayer

One major role of primary visual cortex (V1) in vision is the encoding of the or ientation of lines and contours. The role of the local recurrent network in thes e computations is, however, still a matter of debate. To address this issue, we analyze intracellular recording data of cat V1, which combine measuring the tuni ng of a range of neuronal properties with a precise localization of the recordin g sites in the orientation preference map. For the analysis, we consider a netwo rk model of Hodgkin-Huxley type neurons arranged according to a biologically pla usible two-dimensional topographic orientation preference map. We then systemati cally vary the strength of the recurrent excitation and inhibition relative to t he strength of the afferent input. Each parametrization gives rise to a differen t model instance for which the tuning of model neurons at different locations of the orientation map is compared to the experimentally measured orientation tuni ng of membrane potential, spike output, excitatory, and inhibitory conductances. A quantitative analysis shows that the data provides strong evidence for a netw ork model in which the afferent input is dominated by strong, balanced contribut ions of recurrent excitation and inhibition. This recurrent regime is close to a regime of 'instability', where strong, self-sustained activity of the network o ccurs. The firing rate of neurons in the best-fitting network is particularly se nsitive to small modulations of model parameters, which could be one of the func tional benefits of a network operating in this particular regime.

****

An interior-point stochastic approximation method and an L1-regularized delta ru le

Peter Carbonetto, Mark Schmidt, Nando Freitas

The stochastic approximation method is behind the solution to many important, ac tively-studied problems in machine learning. Despite its far-reaching applicatio n, there is almost no work on applying stochastic approximation to learning prob lems with constraints. The reason for this, we hypothesize, is that no robust, w idely-applicable stochastic approximation method exists for handling such proble ms. We propose that interior-point methods are a natural solution. We establish the stability of a stochastic interior-point approximation method both analytica lly and empirically, and demonstrate its utility by deriving an on-line learning

algorithm that also performs feature selection via L1 regularization.
****
Dynamic visual attention: searching for coding length increments
Xiaodi Hou, Liqing Zhang
A visual attention system should respond placidly when common stimuli are presen ted, while at the same time keep alert to anomalous visual inputs. In this paper , a dynamic visual attention model based on the rarity of features is proposed. We introduce the Incremental Coding Length (ICL) to measure the perspective entr opy gain of each feature. The objective of our model is to maximize the entropy of the sampled visual features. In order to optimize energy consumption, the lim it amount of energy of the system is re-distributed amongst features according t o their Incremental Coding Length. By selecting features with large coding lengt h increments, the computational system can achieve attention selectivity in both static and dynamic scenes. We demonstrate that the proposed model achieves supe rior accuracy in comparison to mainstream approaches in static saliency map gene ration. Moreover, we also show that our model captures several less-reported dyn amic visual search behaviors, such as attentional swing and inhibition of return .
****
Phase transitions for high-dimensional joint support recovery
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-auth ors prior to requesting a name change in the electronic proceedings.
****
Online Metric Learning and Fast Similarity Search
Prateek Jain, Brian Kulis, Inderjit Dhillon, Kristen Grauman
Metric learning algorithms can provide useful distance functions for a variety o f domains, and recent work has shown good accuracy for problems where the learne r can access all distance constraints at once. However, in many real application s, constraints are only available incrementally, thus necessitating methods that can perform online updates to the learned metric. Existing online algorithms of fer bounds on worst-case performance, but typically do not perform well in pract ice as compared to their offline counterparts. We present a new online metric le arning algorithm that updates a learned Mahalanobis metric based on LogDet regul arization and gradient descent. We prove theoretical worst-case performance boun ds, and empirically compare the proposed method against existing online metric l earning algorithms. To further boost the practicality of our approach, we develo p an online locality-sensitive hashing scheme which leads to efficient updates f or approximate similarity search data structures. We demonstrate our algorithm o n multiple datasets and show that it outperforms relevant baselines.
****
Bayesian Network Score Approximation using a Metagraph Kernel
Benjamin Yackley, Eduardo Corona, Terran Lane
Many interesting problems, including Bayesian network structure-search, can be c ast in terms of finding the optimum value of a function over the space of graphs . However, this function is often expensive to compute exactly. We here present a method derived from the study of reproducing-kernel Hilbert spaces which takes advantage of the regular structure of the space of all graphs on a fixed number of nodes to obtain approximations to the desired function quickly and with reas onable accuracy. We then test this method on both a small testing set and a real -world Bayesian network; the results suggest that not only is this method reason ably accurate, but that the BDe score itself varies quadratically over the space of all graphs.
****
Fast Rates for Regularized Objectives
Karthik Sridharan, Shai Shalev-shwartz, Nathan Srebro
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Bayesian Synchronous Grammar Induction
Phil Blunsom, Trevor Cohn, Miles Osborne
We present a novel method for inducing synchronous context free grammars (SCFGs)
 from a corpus of parallel string pairs. SCFGs can model equivalence between str
ings in terms of substitutions, insertions and deletions, and the reordering of
sub-strings. We develop a non-parametric Bayesian model and apply it to a machin
e translation task, using priors to replace the various heuristics commonly used
 in this field. Using a variational Bayes training procedure, we learn the laten
t structure of translation equivalence through the induction of synchronous gram
mar categories for phrasal translations, showing improvements in translation per
formance over previously proposed maximum likelihood models.
****

Tracking Changing Stimuli in Continuous Attractor Neural Networks
K. Wong, Si Wu, Chi Fung
Continuous attractor neural networks (CANNs) are emerging as promising models fo
r describing the encoding of continuous stimuli in neural systems. Due to the tr
anslational invariance of their neuronal interactions, CANNs can hold a continuo
us family of neutrally stable states. In this study, we systematically explore h
ow neutral stability of a CANN facilitates its tracking performance, a capacity
believed to have wide applications in brain functions. We develop a perturbative
 approach that utilizes the dominant movement of the network stationary states i
n the state space. We quantify the distortions of the bump shape during tracking
, and study their effects on the tracking performance. Results are obtained on t
he maximum speed for a moving stimulus to be trackable, and the reaction time to
 catch up an abrupt change in stimulus.
****

Gaussian-process factor analysis for low-dimensional single-trial analysis of ne
ural population activity
Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy
, Maneesh Sahani
We consider the problem of extracting smooth low-dimensional neural trajectories
'' that summarize the activity recorded simultaneously from tens to hundreds of
neurons on individual experimental trials. Beyond the benefit of visualizing the
 high-dimensional noisy spiking activity in a compact denoised form, such trajec
tories can offer insight into the dynamics of the neural circuitry underlying th
e recorded activity. Current methods for extracting neural trajectories involve
a two-stage process: the data are firstdenoised'' by smoothing over time, then a
 static dimensionality reduction technique is applied. We first describe extensi
ons of the two-stage methods that allow the degree of smoothing to be chosen in
a principled way, and account for spiking variability that may vary both across
neurons and across time. We then present a novel method for extracting neural tr
ajectories, Gaussian-process factor analysis (GPFA), which unifies the smoothing
 and dimensionality reduction operations in a common probabilistic framework. We
 applied these methods to the activity of 61 neurons recorded simultaneously in
macaque premotor and motor cortices during reach planning and execution. By adop
ting a goodness-of-fit metric that measures how well the activity of each neuron
 can be predicted by all other recorded neurons, we found that GPFA provided a b
etter characterization of the population activity than the two-stage methods. Fr
om the extracted single-trial neural trajectories, we directly observed a conver
gence in neural state during motor planning, an effect suggestive of attractor d
ynamics that was shown indirectly by previous studies.
****

Kernel Measures of Independence for non-iid Data
Xinhua Zhang, Le Song, Arthur Gretton, Alex Smola
Many machine learning algorithms can be formulated in the framework of statistic
al independence such as the Hilbert Schmidt Independence Criterion. In this pape
r, we extend this criterion to deal with with structured and interdependent obse

rvations. This is achieved by modeling the structures using undirected graphical models and comparing the Hilbert space embeddings of distributions. We apply this new criterion to independent component analysis and sequence clustering.

****

Regularized Policy Iteration

Amir Farahmand, Mohammad Ghavamzadeh, Shie Mannor, Csaba Szepesvári

In this paper we consider approximate policy-iteration-based reinforcement learning algorithms. In order to implement a flexible function approximation scheme we propose the use of non-parametric methods with regularization, providing a convenient way to control the complexity of the function approximator. We propose two novel regularized policy iteration algorithms by adding L2-regularization to two widely-used policy evaluation methods: Bellman residual minimization (BRM) and least-squares temporal difference learning (LSTD). We derive efficient implementation for our algorithms when the approximate value-functions belong to a reproducing kernel Hilbert space. We also provide finite-sample performance bounds for our algorithms and show that they are able to achieve optimal rates of convergence under the studied conditions.

****

Dimensionality Reduction for Data in Multiple Feature Representations

Yen-yu Lin, Tyng-luh Liu, Chiou-shann Fuh

In solving complex visual learning tasks, adopting multiple descriptors to more precisely characterize the data has been a feasible way for improving performance. These representations are typically high dimensional and assume diverse forms. Thus finding a way to transform them into a unified space of lower dimension generally facilitates the underlying tasks, such as object recognition or clustering. We describe an approach that incorporates multiple kernel learning with dimensionality reduction (MKL-DR). While the proposed framework is flexible in simultaneously tackling data in various feature representations, the formulation itself is general in that it is established upon graph embedding. It follows that any dimensionality reduction techniques explainable by graph embedding can be generalized by our method to consider data in multiple feature representations.

****

Continuously-adaptive discretization for message-passing algorithms

Michael Isard, John MacCormick, Kannan Achan

Continuously-Adaptive Discretization for Message-Passing (CAD-MP) is a new message-passing algorithm employing adaptive discretization. Most previous message-passing algorithms approximated arbitrary continuous probability distributions using either: a family of continuous distributions such as the exponential family; a particle-set of discrete samples; or a fixed, uniform discretization. In contrast, CAD-MP uses a discretization that is (i) non-uniform, and (ii) adaptive. The non-uniformity allows CAD-MP to localize interesting features (such as sharp peaks) in the marginal belief distributions with time complexity that scales logarithmically with precision, as opposed to uniform discretization which scales at best linearly. We give a principled method for altering the non-uniform discretization according to information-based measures. CAD-MP is shown in experiments on simulated data to estimate marginal beliefs much more precisely than competing approaches for the same computational expense.

****

On Computational Power and the Order-Chaos Phase Transition in Reservoir Computing

Benjamin Schrauwen, Lars Buesing, Robert Legenstein

Randomly connected recurrent neural circuits have proven t o be very powerful models for online computations when a trained memoryless re adout function is appended. Such Reservoir Computing (RC) systems are commonly used in two flavors: with analog or binary (spiking) neurons in the recur rent circuits. Previous work showed a fundamental difference between these two incarnations of the RC idea. The performance of a RC system built from binary neuron s seems to depend strongly on the network connectivity structure. In network s of analog neurons such dependency has not been observed. In this article we investigate this apparent dichotomy in terms of the in-degree of the circuit nodes. Our analyses based amongst

others on the Lyapunov exponent reveal that the phase transition between ordered and chaotic network behavior of binary circuits qualitatively differs from the one in analog circuits. This explains the observed decreased computational performance of binary circuits of high node in-degree. Furthermore, a novel mean-field predictor for computational performance is introduced and shown to accurately predict the numerically obtained results.

****

Mind the Duality Gap: Logarithmic regret algorithms for online optimization

Shai Shalev-shwartz, Sham M. Kakade

We describe a primal-dual framework for the design and analysis of online strongly convex optimization algorithms. Our framework yields the tightest known logarithmic regret bounds for Follow-The-Leader and for the gradient descent algorithm proposed in HazanKaKaAg06. We then show that one can interpolate between these two extreme cases. In particular, we derive a new algorithm that shares the computational simplicity of gradient descent but achieves lower regret in many practical situations. Finally, we further extend our framework for generalized strongly convex functions.

****

Hierarchical Semi-Markov Conditional Random Fields for Recursive Sequential Data

Tran Truyen, Dinh Phung, Hung Bui, Svetha Venkatesh

Inspired by the hierarchical hidden Markov models (HHMM), we present the hierarchical semi-Markov conditional random field (HSCRF), a generalisation of embedded undirected Markov chains to model complex hierarchical, nested Markov processes. It is parameterised in a discriminative framework and has polynomial time algorithms for learning and inference. Importantly, we develop efficient algorithms for learning and constrained inference in a partially-supervised setting, which is important issue in practice where labels can only be obtained sparsely. We demonstrate the HSCRF in two applications: (i) recognising human activities of daily living (ADLs) from indoor surveillance cameras, and (ii) noun-phrase chunking. We show that the HSCRF is capable of learning rich hierarchical models with reasonable accuracy in both fully and partially observed data cases.

****

A spatially varying two-sample recombinant coalescent, with applications to HIV escape response

Alexander Braunstein, Zhi Wei, Shane Jensen, Jon Mcauliffe

Statistical evolutionary models provide an important mechanism for describing and understanding the escape response of a viral population under a particular therapy. We present a new hierarchical model that incorporates spatially varying mutation and recombination rates at the nucleotide level. It also maintains sep- arate parameters for treatment and control groups, which allows us to estimate treatment effects explicitly. We use the model to investigate the sequence evolu- tion of HIV populations exposed to a recently developed antisense gene therapy, as well as a more conventional drug therapy. The detection of biologically rele- vant and plausible signals in both therapy studies demonstrates the effectiveness of the method.

****

Measures of Clustering Quality: A Working Set of Axioms for Clustering

Shai Ben-David, Margareta Ackerman

Aiming towards the development of a general clustering theory, we discuss abstract axiomatization for clustering. In this respect, we follow up on the work of Kelinberg, (Kleinberg) that showed an impossibility result for such axiomatization. We argue that an impossibility result is not an inherent feature of clustering, but rather, to a large extent, it is an artifact of the specific formalism used in Kleinberg. As opposed to previous work focusing on clustering functions, we propose to address clustering quality measures as the primitive object to be axiomatized. We show that principles like those formulated in Kleinberg's axioms can be readily expressed in the latter framework without leading to inconsistency. A clustering-quality measure is a function that, given a data set and its partition into clusters, returns a non-negative real number representing how strong ' orconclusive' the clustering is. We analyze what clustering-quality measures s

hould look like and introduce a set of requirements (`axioms') that express thes e requirement and extend the translation of Kleinberg's axioms to our framework. We propose several natural clustering quality measures, all satisfying the prop osed axioms. In addition, we show that the proposed clustering quality can be co mputed in polynomial time.

****

Structure Learning in Human Sequential Decision-Making

Daniel Acuna, Paul R. Schrater

We use graphical models and structure learning to explore how people learn polic ies in sequential decision making tasks. Studies of sequential decision-making i n humans frequently find suboptimal performance relative to an ideal actor that knows the graph model that generates reward in the environment. We argue that th e learning problem humans face also involves learning the graph structure for re ward generation in the environment. We formulate the structure learning problem using mixtures of reward models, and solve the optimal action selection problem using Bayesian Reinforcement Learning. We show that structure learning in one an d two armed bandit problems produces many of the qualitative behaviors deemed su boptimal in previous studies. Our argument is supported by the results of experi ments that demonstrate humans rapidly learn and exploit new reward structure.

****

Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Perf ormance

Jeremy Hill, Jason Farquhar, Suzanna Martens, Felix Biessmann, Bernhard Schölkop f

From an information-theoretic perspective, a noisy transmission system such as a visual Brain-Computer Interface (BCI) speller could benefit from the use of err or-correcting codes. However, optimizing the code solely according to the maxima l minimum-Hamming-distance criterion tends to lead to an overall increase in tar get frequency of target stimuli, and hence a significantly reduced average targe t-to-target interval (TTI), leading to difficulties in classifying the individua l event-related potentials (ERPs) due to overlap and refractory effects. Clearly any change to the stimulus setup must also respect the possible psychophysiolog ical consequences. Here we report new EEG data from experiments in which we expl ore stimulus types and codebooks in a within-subject design, finding an interact ion between the two factors. Our data demonstrate that the traditional, row-colu mn code has particular spatial properties that lead to better performance than o ne would expect from its TTIs and Hamming-distances alone, but nonetheless error -correcting codes can improve performance provided the right stimulus type is us ed.

****

Supervised Dictionary Learning

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, Francis Bach

It is now well established that sparse signal models are well suited to restorat ion tasks and can effectively be learned from audio, image, and video data. Rece nt research has been aimed at learning discriminative sparse models instead of p urely reconstructive ones. This paper proposes a new step in that direction with a novel sparse representation for signals belonging to different classes in ter ms of a shared dictionary and multiple decision functions. It is shown that the linear variant of the model admits a simple probabilistic interpretation, and th at its most general variant also admits a simple interpretation in terms of kern els. An optimization framework for learning all the components of the proposed m odel is presented, along with experiments on standard handwritten digit and text ure classification tasks.

****

Natural Image Denoising with Convolutional Networks

Viren Jain, Sebastian Seung

We present an approach to low-level vision that combines two main ideas: the use of convolutional networks as an image processing architecture and an unsupervis ed learning procedure that synthesizes training samples from specific noise mode ls. We demonstrate this approach on the challenging problem of natural image den

oising. Using a test set with a hundred natural images, we find that convolution al networks provide comparable and in some cases superior performance to state o f the art wavelet and Markov random field (MRF) methods. Moreover, we find that a convolutional network offers similar performance in the blind denoising settin g as compared to other techniques in the non-blind setting. We also show how con volutional networks are mathematically related to MRF approaches by presenting a mean field theory for an MRF specially designed for image denoising. Although t hese approaches are related, convolutional networks avoid computational difficul ties in MRF approaches that arise from probabilistic learning and inference. Thi s makes it possible to learn image processing architectures that have a high deg ree of representational power (we train models with over 15,000 parameters), but whose computational expense is significantly less than that associated with inf erence in MRF approaches with even hundreds of parameters.

****

Optimal Response Initiation: Why Recent Experience Matters

Matt Jones, Sachiko Kinoshita, Michael C. Mozer

In most cognitive and motor tasks, speed-accuracy tradeoffs are observed: Indivi duals can respond slowly and accurately, or quickly yet be prone to errors. Cont rol mechanisms governing the initiation of behavioral responses are sensitive no t only to task instructions and the stimulus being processed, but also to the re cent stimulus history. When stimuli can be characterized on an easy-hard dimensi on (e.g., word frequency in a naming task), items preceded by easy trials are re sponded to more quickly, and with more errors, than items preceded by hard trial s. We propose a rationally motivated mathematical model of this sequential adapt ation of control, based on a diffusion model of the decision process in which di fficulty corresponds to the drift rate for the correct response. The model assum es that responding is based on the posterior distribution over which response is correct, conditioned on the accumulated evidence. We derive this posterior as a function of the drift rate, and show that higher estimates of the drift rate le ad to (normatively) faster responding. Trial-by-trial tracking of difficulty thu s leads to sequential effects in speed and accuracy. Simulations show the model explains a variety of phenomena in human speeded decision making. We argue this passive statistical mechanism provides a more elegant and parsimonious account t han extant theories based on elaborate control structures.

****

Bayesian Experimental Design of Magnetic Resonance Imaging Sequences

Hannes Nickisch, Rolf Pohmann, Bernhard Schölkopf, Matthias Seeger

We show how improved sequences for magnetic resonance imaging can be found throu gh automated optimization of Bayesian design scores. Combining recent advances i n approximate Bayesian inference and natural image statistics with high-performa nce numerical computation, we propose the first scalable Bayesian experimental d esign framework for this problem of high relevance to clinical and brain researc h. Our solution requires approximate inference for dense, non-Gaussian models on a scale seldom addressed before. We propose a novel scalable variational infere nce algorithm, and show how powerful methods of numerical mathematics can be mod ified to compute primitives in our framework. Our approach is evaluated on a rea listic setup with raw data from a 3T MR scanner.

****

Temporal Dynamics of Cognitive Control

Jeremy Reynolds, Michael C. Mozer

Cognitive control refers to the flexible deployment of memory and attention in r esponse to task demands and current goals. Control is often studied experimental ly by presenting sequences of stimuli, some demanding a response, and others mod ulating the stimulus-response mapping. In these tasks, participants must maintai n information about the current stimulus-response mapping in working memory. Pro minent theories of cognitive control use recurrent neural nets to implement work ing memory, and optimize memory utilization via reinforcement learning. We prese nt a novel perspective on cognitive control in which working memory representati ons are intrinsically probabilistic, and control operations that maintain and up date working memory are dynamically determined via probabilistic inference. We s

how that our model provides a parsimonious account of behavioral and neuroimagin
g data, and suggest that it offers an elegant conceptualization of control in wh
ich behavior can be cast as optimal, subject to limitations on learning and the
rate of information processing. Moreover, our model provides insight into how ta
sk instructions can be directly translated into appropriate behavior and then ef
ficiently refined with subsequent task experience.
****

Overlaying classifiers: a practical approach for optimal ranking
Stéphan Clémençcon, Nicolas Vayatis
ROC curves are one of the most widely used displays to evaluate performance of s
coring functions. In the paper, we propose a statistical method for directly opt
imizing the ROC curve. The target is known to be the regression function up to a
n increasing transformation and this boils down to recovering the level sets of
the latter. We propose to use classifiers obtained by empirical risk minimizatio
n of a weighted classification error and then to construct a scoring rule by ove
rlaying these classifiers. We show the consistency and rate of convergence to th
e optimal ROC curve of this procedure in terms of supremum norm and also, as a b
yproduct of the analysis, we derive an empirical estimate of the optimal ROC cur
ve.
****

Model selection and velocity estimation using novel priors for motion patterns
Shuang Wu, Hongjing Lu, Alan L. Yuille
Psychophysical experiments show that humans are better at perceiving rotation an
d expansion than translation. These findings are inconsistent with standard mode
ls of motion integration which predict best performance for translation [6]. To
explain this discrepancy, our theory formulates motion perception at two levels
of inference: we first perform model selection between the competing models (e.g
. translation, rotation, and expansion) and then estimate the velocity using the
 selected model. We define novel prior models for smooth rotation and expansion
using techniques similar to those in the slow-and-smooth model [17] (e.g. Green
functions of differential operators). The theory gives good agreement with the t
rends observed in human experiments.
****

Estimation of Information Theoretic Measures for Continuous Random Variables
Fernando Pérez-Cruz
We analyze the estimation of information theoretic measures of continuous random
 variables such as: differential entropy, mutual information or Kullback-Leibler
 divergence. The objective of this paper is two-fold. First, we prove that the i
nformation theoretic measure estimates using the k-nearest-neighbor density esti
mation with fixed k converge almost surely, even though the k-nearest-neighbor d
ensity estimation with fixed k does not converge to its true measure. Second, we
 show that the information theoretic measure estimates do not converge for k gro
wing linearly with the number of samples. Nevertheless, these nonconvergent esti
mates can be used for solving the two-sample problem and assessing if two random
 variables are independent. We show that the two-sample and independence tests b
ased on these nonconvergent estimates compare favorably with the maximum mean di
screpancy test and the Hilbert Schmidt independence criterion, respectively.
****

On the Reliability of Clustering Stability in the Large Sample Regime
Ohad Shamir, Naftali Tishby
Clustering stability is an increasingly popular family of methods for performing
 model selection in data clustering. The basic idea is that the chosen model sho
uld be stable under perturbation or resampling of the data. Despite being reason
ably effective in practice, these methods are not well understood theoretically,
 and present some difficulties. In particular, when the data is assumed to be sa
mpled from an underlying distribution, the solutions returned by the clustering
algorithm will usually become more and more stable as the sample size increases.
 This raises a potentially serious practical difficulty with these methods, beca
use it means there might be some hard-to-compute sample size, beyond which clust
ering stability estimators 'break down' and become unreliable in detecting the m

ost stable model. Namely, all models will be relatively stable, with differences in their stability measures depending mostly on random and meaningless sampling artifacts. In this paper, we provide a set of general sufficient conditions, which ensure the reliability of clustering stability estimators in the large sample regime. In contrast to previous work, which concentrated on specific toy distributions or specific idealized clustering frameworks, here we make no such assumptions. We then exemplify how these conditions apply to several important families of clustering algorithms, such as maximum likelihood clustering, certain types of kernel clustering, and centroid-based clustering with any Bregman divergence. In addition, we explicitly derive the non-trivial asymptotic behavior of these estimators, for any framework satisfying our conditions. This can help us understand what is considered a 'stable' model by these estimators, at least for large enough samples.

****

Using matrices to model symbolic relationship

Ilya Sutskever, Geoffrey E. Hinton

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.

****

Psychiatry: Insights into depression through normative decision-making models

Quentin Huys, Joshua Vogelstein, Peter Dayan

Decision making lies at the very heart of many psychiatric diseases. It is also a central theoretical concern in a wide variety of ■elds and has undergone detailed, in-depth, analyses. We take as an example Major Depressive Disorder (MDD), applying insights from a Bayesian reinforcement learning framework. We focus on anhedonia and helplessness. Helplessness—a core element in the conceptual- izations of MDD that has lead to major advances in its treatment, pharmacolog- ical and neurobiological understanding—is formalized as a simple prior over the outcome entropy of actions in uncertain environments. Anhedonia, which is an equally fundamental aspect of the disease, is related to the effective reward size. These formulations allow for the design of speci■c tasks to measure anhedonia and helplessness behaviorally. We show that these behavioral measures capture explicit, questionnaire-based cognitions. We also provide evidence that these tasks may allow classi■cation of subjects into healthy and MDD groups based purely on a behavioural measure and avoiding any verbal reports.

****

Characteristic Kernels on Groups and Semigroups

Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Bharath K. Sriperumbudur

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

****

Multi-label Multiple Kernel Learning

Shuiwang Ji, Liang Sun, Rong Jin, Jieping Ye

We present a multi-label multiple kernel learning (MKL) formulation, in which the data are embedded into a low-dimensional space directed by the instance-label correlations encoded into a hypergraph. We formulate the problem in the kernel-induced feature space and propose to learn the kernel matrix as a linear combination of a given collection of kernel matrices in the MKL framework. The proposed learning formulation leads to a non-smooth min-max problem, and it can be cast into a semi-infinite linear program (SILP). We further propose an approximate formulation with a guaranteed error bound which involves an unconstrained and convex optimization problem. In addition, we show that the objective function of the approximate formulation is continuously differentiable with Lipschitz gradient, and hence existing methods can be employed to compute the optimal solution efficiently. We apply the proposed formulation to the automated annotation of Drosophila gene expression pattern images, and promising results have been reported in

comparison with representative algorithms.
****
Bio-inspired Real Time Sensory Map Realignment in a Robotic Barn Owl
Juan Huo, Zhijun Yang, Alan Murray
The visual and auditory map alignment in the Superior Colliculus (SC) of barn ow l is important for its accurate localization for prey behavior. Prism learning o r Blindness may interfere this alignment and cause loss of the capability of acc urate prey. However, juvenile barn owl could recover its sensory map alignment b y shifting its auditory map. The adaptation of this map alignment is believed ba sed on activity dependent axon developing in Inferior Colliculus (IC). A model i s built to explore this mechanism. In this model, axon growing process is instru cted by an inhibitory network in SC while the strength of the inhibition adjuste d by Spike Timing Dependent Plasticity (STDP). We test and analyze this mechanis m by application of the neural structures involved in spatial localization in a robotic system.
****
Spectral Hashing
Yair Weiss, Antonio Torralba, Rob Fergus
Semantic hashing seeks compact binary codes of datapoints so that the Hamming di stance between codewords correlates with semantic similarity. Hinton et al. used a clever implementation of autoencoders to find such codes. In this paper, we s how that the problem of finding a best code for a given dataset is closely relat ed to the problem of graph partitioning and can be shown to be NP hard. By relax ing the original problem, we obtain a spectral method whose solutions are simply a subset of thresh- olded eigenvectors of the graph Laplacian. By utilizing rec ent results on convergence of graph Laplacian eigenvectors to the Laplace-Beltra mi eigen- functions of manifolds, we show how to efficiently calculate the code of a novel datapoint. Taken together, both learning the code and applying it to a novel point are extremely simple. Our experiments show that our codes signific antly outperform the state-of-the art.
****
Multi-resolution Exploration in Continuous Spaces
Ali Nouri, Michael Littman
The essence of exploration is acting to try to decrease uncertainty. We propose a new methodology for representing uncertainty in continuous-state control probl ems. Our approach, multi-resolution exploration (MRE), uses a hierarchical mappi ng to identify regions of the state space that would benefit from additional sam ples. We demonstrate MRE's broad utility by using it to speed up learning in a p rototypical model-based and value-based reinforcement-learning method. Empirical results show that MRE improves upon state-of-the-art exploration approaches.
****
A computational model of hippocampal function in trace conditioning
Elliot Ludvig, Richard S. Sutton, Eric Verbeek, E. Kehoe
We present a new reinforcement-learning model for the role of the hippocampus in classical conditioning, focusing on the differences between trace and delay con ditioning. In the model, all stimuli are represented both as unindividuated whol es and as a series of temporal elements with varying delays. These two stimulus representations interact, producing different patterns of learning in trace and delay conditioning. The model proposes that hippocampal lesions eliminate long-l atency temporal elements, but preserve short-latency temporal elements. For trac e conditioning, with no contiguity between stimulus and reward, these long-laten cy temporal elements are vital to learning adaptively timed responses. For delay conditioning, in contrast, the continued presence of the stimulus supports cond itioned responding, and the short-latency elements suppress responding early in the stimulus. In accord with the empirical data, simulated hippocampal damage im pairs trace conditioning, but not delay conditioning, at medium-length intervals . With longer intervals, learning is impaired in both procedures, and, with shor ter intervals, in neither. In addition, the model makes novel predictions about the response topography with extended stimuli or post-training lesions. These re sults demonstrate how temporal contiguity, as in delay conditioning, changes the

timing problem faced by animals, rendering it both easier and less susceptible to disruption by hippocampal lesions.
****

Online Prediction on Large Diameter Graphs
Mark Herbster, Guy Lever, Massimiliano Pontil
Current on-line learning algorithms for predicting the labelling of a graph have an important limitation in the case of large diameter graphs; the number of mistakes made by such algorithms may be proportional to the square root of the number of vertices, even when tackling simple problems. We overcome this problem with an efficient algorithm which achieves a logarithmic mistake bound. Furthermore, current algorithms are optimised for data which exhibits cluster-structure; we give an additional algorithm which performs well locally in the presence of cluster structure and on large diameter graphs.
****

Sparse probabilistic projections
Cédric Archambeau, Francis Bach
We present a generative model for performing sparse probabilistic projections, which includes sparse principal component analysis and sparse canonical correlation analysis as special cases. Sparsity is enforced by means of automatic relevance determination or by imposing appropriate prior distributions, such as generalised hyperbolic distributions. We derive a variational Expectation-Maximisation algorithm for the estimation of the hyperparameters and show that our novel probabilistic approach compares favourably to existing techniques. We illustrate how the proposed method can be applied in the context of cryptoanalysis as a pre-processing tool for the construction of template attacks.
****

Sparsity of SVMs that use the epsilon-insensitive loss
Ingo Steinwart, Andreas Christmann
In this paper lower and upper bounds for the number of support vectors are derived for support vector machines (SVMs) based on the epsilon-insensitive loss function. It turns out that these bounds are asymptotically tight under mild assumptions on the data generating distribution. Finally, we briefly discuss a trade-off in epsilon between sparsity and accuracy if the SVM is used to estimate the conditional median.
****

Automatic online tuning for fast Gaussian summation
Vlad Morariu, Balaji Srinivasan, Vikas C. Raykar, Ramani Duraiswami, Larry S. Davis
Many machine learning algorithms require the summation of Gaussian kernel functions, an expensive operation if implemented straightforwardly. Several methods have been proposed to reduce the computational complexity of evaluating such sums, including tree and analysis based methods. These achieve varying speedups depending on the bandwidth, dimension, and prescribed error, making the choice between methods difficult for machine learning tasks. We provide an algorithm that combines tree methods with the Improved Fast Gauss Transform (IFGT). As originally proposed the IFGT suffers from two problems: (1) the Taylor series expansion does not perform well for very low bandwidths, and (2) parameter selection is not trivial and can drastically affect performance and ease of use. We address the first problem by employing a tree data structure, resulting in four evaluation methods whose performance varies based on the distribution of sources and targets and input parameters such as desired accuracy and bandwidth. To solve the second problem, we present an online tuning approach that results in a black box method that automatically chooses the evaluation method and its parameters to yield the best performance for the input data, desired accuracy, and bandwidth. In addition, the new IFGT parameter selection approach allows for tighter error bounds. Our approach chooses the fastest method at negligible additional cost, and has superior performance in comparisons with previous approaches.
****

Reducing statistical dependencies in natural signals using radial Gaussianization

Siwei Lyu, Eero Simoncelli
We consider the problem of efficiently encoding a signal by transforming it to a new representation whose components are statistically independent. A widely studied linear solution, independent components analysis (ICA), exists for the case when the signal is generated as a linear transformation of independent non- Gaussian sources. Here, we examine a complementary case, in which the source is non-Gaussian but elliptically symmetric. In this case, no linear transform suffices to properly decompose the signal into independent components, but we show that a simple nonlinear transformation, which we call radial Gaussianization (RG), is able to remove all dependencies. We then demonstrate this methodology in the context of natural signal statistics. We first show that the joint distributions of bandpass filter responses, for both sound and images, are better described as elliptical than linearly transformed independent sources. Consistent with this, we demonstrate that the reduction in dependency achieved by applying RG to either pairs or blocks of bandpass filter responses is significantly greater than that achieved by PCA or ICA.
****

A Transductive Bound for the Voted Classifier with an Application to Semi-supervised Learning
Massih R. Amini, Nicolas Usunier, François Laviolette
In this paper we present two transductive bounds on the risk of the majority vote estimated over partially labeled training sets. Our first bound is tight when the additional unlabeled training data are used in the cases where the voted classifier makes its errors on low margin observations and where the errors of the associated Gibbs classifier can accurately be estimated. In semi-supervised learning, considering the margin as an indicator of confidence constitutes the working hypothesis of algorithms which search the decision boundary on low density regions. In this case, we propose a second bound on the joint probability that the voted classifier makes an error over an example having its margin over a fixed threshold. As an application we are interested on self-learning algorithms which assign iteratively pseudo-labels to unlabeled training examples having margin above a threshold obtained from this bound. Empirical results on different datasets show the effectiveness of our approach compared to the same algorithm and the TSVM in which the threshold is fixed manually.
****

Nonrigid Structure from Motion in Trajectory Space
Ijaz Akhter, Yaser Sheikh, Sohaib Khan, Takeo Kanade
Existing approaches to nonrigid structure from motion assume that the instantaneous 3D shape of a deforming object is a linear combination of basis shapes, which have to be estimated anew for each video sequence. In contrast, we propose that the evolving 3D structure be described by a linear combination of basis trajectories. The principal advantage of this lateral approach is that we do not need to estimate any basis vectors during computation. Instead, we show that generic bases over trajectories, such as the Discrete Cosine Transform (DCT) bases, can be used to effectively describe most real motions. This results in a significant reduction in unknowns, and corresponding stability, in estimation. We report empirical performance, quantitatively using motion capture data and qualitatively on several video sequences exhibiting nonrigid motions including piece-wise rigid motion, articulated motion, partially nonrigid motion (such as a facial expression), and highly nonrigid motion (such as a person dancing).
****

Transfer Learning by Distribution Matching for Targeted Advertising
Steffen Bickel, Christoph Sawade, Tobias Scheffer
We address the problem of learning classifiers for several related tasks that may differ in their joint distribution of input and output variables. For each task, small - possibly even empty - labeled samples and large unlabeled samples are available. While the unlabeled samples reflect the target distribution, the labeled samples may be biased. We derive a solution that produces resampling weights which match the pool of all examples to the target distribution of any given task. Our work is motivated by the problem of predicting sociodemographic feature

s for users of web portals, based on the content which they have accessed. Here, questionnaires offered to a small portion of each portal's users produce biased samples. Transfer learning enables us to make predictions even for new portals with few or no training data and improves the overall prediction accuracy.
****
A Convergent $O(n)$ Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation
Richard S. Sutton, Hamid Maei, Csaba Szepesvári
We introduce the first temporal-difference learning algorithm that is stable with linear function approximation and off-policy training, for any finite Markov decision process, target policy, and exciting behavior policy, and whose complexity scales linearly in the number of parameters. We consider an i.i.d.\ policy-evaluation setting in which the data need not come from on-policy experience. The gradient temporal-difference (GTD) algorithm estimates the expected update vector of the TD(0) algorithm and performs stochastic gradient descent on its L_2 norm. Our analysis proves that its expected update is in the direction of the gradient, assuring convergence under the usual stochastic approximation conditions to the same least-squares solution as found by the LSTD, but without its quadratic computational complexity. GTD is online and incremental, and does not involve multiplying by products of likelihood ratios as in importance-sampling methods.
****
Estimating Robust Query Models with Convex Optimization
Kevyn Collins-thompson
Query expansion is a long-studied approach for improving retrieval effectiveness by enhancing the userâ■■s original query with additional related terms. Current algorithms for automatic query expansion have been shown to consistently improve retrieval accuracy on average, but are highly unstable and have bad worst-case performance for individual queries. We introduce a novel risk framework that formulates query model estimation as a constrained metric labeling problem on a graph of term relations. Themodel combines assignment costs based on a baseline feedback algorithm, edge weights based on term similarity, and simple constraints to enforce aspect balance, aspect coverage, and term centrality. Results across multiple standard test collections show consistent and dramatic reductions in the number and magnitude of expansion failures, while retaining the strong positive gains of the baseline algorithm.
****
Learning Taxonomies by Dependence Maximization
Matthew Blaschko, Arthur Gretton
We introduce a family of unsupervised algorithms, numerical taxonomy clustering, to simultaneously cluster data, and to learn a taxonomy that encodes the relationship between the clusters. The algorithms work by maximizing the dependence between the taxonomy and the original data. The resulting taxonomy is a more informative visualization of complex data than simple clustering; in addition, taking into account the relations between different clusters is shown to substantially improve the quality of the clustering, when compared with state-of-the-art algorithms in the literature (both spectral clustering and a previous dependence maximization approach). We demonstrate our algorithm on image and text data.
****
Finding Latent Causes in Causal Networks: an Efficient Approach Based on Markov Blankets
Jean-philippe Pellet, André Elisseeff
Causal structure-discovery techniques usually assume that all causes of more than one variable are observed. This is the so-called causal sufficiency assumption. In practice, it is untestable, and often violated. In this paper, we present an efficient causal structure-learning algorithm, suited for causally insufficient data. Similar to algorithms such as IC* and FCI, the proposed approach drops the causal sufficiency assumption and learns a structure that indicates (potential) latent causes for pairs of observed variables. Assuming a constant local density of the data-generating graph, our algorithm makes a quadratic number of conditional-independence tests w.r.t. the number of variables. We show with experime

nts that our algorithm is comparable to the state-of-the-art FCI algorithm in ac curacy, while being several orders of magnitude faster on large problems. We con clude that MBCS* makes a new range of causally insufficient problems computation ally tractable.
****

Goal-directed decision making in prefrontal cortex: a computational framework
Matthew Botvinick, James An
Research in animal learning and behavioral neuroscience has distinguished betwee n two forms of action control: a habit-based form, which relies on stored action values, and a goal-directed form, which forecasts and compares action outcomes based on a model of the environment. While habit-based control has been the subj ect of extensive computational research, the computational principles underlying goal-directed control in animals have so far received less attention. In the pr esent paper, we advance a computational framework for goal-directed control in a nimals and humans. We take three empirically motivated points as founding premis es: (1) Neurons in dorsolateral prefrontal cortex represent action policies, (2) Neurons in orbitofrontal cortex represent rewards, and (3) Neural computation, across domains, can be appropriately understood as performing structured probabi listic inference. On a purely computational level, the resulting account relates closely to previous work using Bayesian inference to solve Markov decision prob lems, but extends this work by introducing a new algorithm, which provably conve rges on optimal plans. On a cognitive and neuroscientific level, the theory prov ides a unifying framework for several different forms of goal-directed action se lection, placing emphasis on a novel form, within which orbitofrontal reward rep resentations directly drive policy selection.
****

Localized Sliced Inverse Regression
Qiang Wu, Sayan Mukherjee, Feng Liang
We developed localized sliced inverse regression for supervised dimension reduct ion. It has the advantages of preventing degeneracy, increasing estimation accur acy, and automatic subclass discovery in classification problems. A semisupervis ed version is proposed for the use of unlabeled data. The utility is illustrated on simulated as well as real data sets.
****

Near-optimal Regret Bounds for Reinforcement Learning
Peter Auer, Thomas Jaksch, Ronald Ortner
For undiscounted reinforcement learning in Markov decision processes (MDPs) we c onsider the total regret of a learning algorithm with respect to an optimal poli cy. In order to describe the transition structure of an MDP we propose a new par ameter: An MDP has diameter D if for any pair of states s1,s2 there is a policy which moves from s1 to s2 in at most D steps (on average). We present a reinforc ement learning algorithm with total regret O(DSAT) after T steps for any unknown MDP with S states, A actions per state, and diameter D. This bound holds with h igh probability. We also present a corresponding lower bound of Omega(DSAT) on t he total regret of any learning algorithm. Both bounds demonstrate the utility o f the diameter as structural parameter of the MDP.
****

Non-parametric Regression Between Manifolds
Florian Steinke, Matthias Hein
This paper discusses non-parametric regression between Riemannian manifolds. Thi s learning problem arises frequently in many application areas ranging from sign al processing, computer vision, over robotics to computer graphics. We present a new algorithmic scheme for the solution of this general learning problem based on regularized empirical risk minimization. The regularization functional takes into account the geometry of input and output manifold, and we show that it impl ements a prior which is particularly natural. Moreover, we demonstrate that our algorithm performs well in a difficult surface registration problem.
****

Unsupervised Learning of Visual Sense Models for Polysemous Words
Kate Saenko, Trevor Darrell

Polysemy is a problem for methods that exploit image search engines to build obj
ect category models. Existing unsupervised approaches do not take word sense int
o consideration. We propose a new method that uses a dictionary to learn models
of visual word sense from a large collection of unlabeled web data. The use of L
DA to discover a latent sense space makes the model robust despite the very limi
ted nature of dictionary definitions. The definitions are used to learn a distri
bution in the latent space that best represents a sense. The algorithm then uses
 the text surrounding image links to retrieve images with high probability of a
particular dictionary sense. An object classifier is trained on the resulting se
nse-specific images. We evaluate our method on a dataset obtained by searching t
he web for polysemous words. Category classification experiments show that our d
ictionary-based approach outperforms baseline methods.
****

Extended Grassmann Kernels for Subspace-Based Learning
Jihun Hamm, Daniel Lee
Subspace-based learning problems involve data whose elements are linear subspace
s of a vector space. To handle such data structures, Grassmann kernels have been
 proposed and used previously. In this paper, we analyze the relationship betwee
n Grassmann kernels and probabilistic similarity measures. Firstly, we show that
 the KL distance in the limit yields the Projection kernel on the Grassmann mani
fold, whereas the Bhattacharyya kernel becomes trivial in the limit and is subop
timal for subspace-based problems. Secondly, based on our analysis of the KL dis
tance, we propose extensions of the Projection kernel which can be extended to t
he set of affine as well as scaled subspaces. We demonstrate the advantages of t
hese extended kernels for classification and recognition tasks with Support Vect
or Machines and Kernel Discriminant Analysis using synthetic and real image data
bases.
****

Implicit Mixtures of Restricted Boltzmann Machines
Vinod Nair, Geoffrey E. Hinton
We present a mixture model whose components are Restricted Boltzmann Machines (R
BMs). This possibility has not been considered before because computing the part
ition function of an RBM is intractable, which appears to make learning a mixtur
e of RBMs intractable as well. Surprisingly, when formulated as a third-order Bo
ltzmann machine, such a mixture model can be learned tractably using contrastive
 divergence. The energy function of the model captures three-way interactions am
ong visible units, hidden units, and a single hidden multinomial unit that repre
sents the cluster labels. The distinguishing feature of this model is that, unli
ke other mixture models, the mixing proportions are not explicitly parameterized
. Instead, they are defined implicitly via the energy function and depend on all
 the parameters in the model. We present results for the MNIST and NORB datasets
 showing that the implicit mixture of RBMs learns clusters that reflect the clas
s structure in the data.
****

Self-organization using synaptic plasticity
Vicençc Gómez, Andreas Kaltenbrunner, Vicente López, Hilbert Kappen
Large networks of spiking neurons show abrupt changes in their collective dynami
cs resembling phase transitions studied in statistical physics. An example of th
is phenomenon is the transition from irregular, noise-driven dynamics to regular
, self-sustained behavior observed in networks of integrate-and-fire neurons as
the interaction strength between the neurons increases. In this work we show how
 a network of spiking neurons is able to self-organize towards a critical state
for which the range of possible inter-spike-intervals (dynamic range) is maximiz
ed. Self-organization occurs via synaptic dynamics that we analytically derive.
The resulting plasticity rule is defined locally so that global homeostasis near
 the critical state is achieved by local regulation of individual synapses.
****

Interpreting the neural code with Formal Concept Analysis
Dominik Endres, Peter Foldiak
We propose a novel application of Formal Concept Analysis (FCA) to neural decodi

ng: instead of just trying to figure out which stimulus was presented, we demons trate how to explore the semantic relationships between the neural representatio n of large sets of stimuli. FCA provides a way of displaying and interpreting su ch relationships via concept lattices. We explore the effects of neural code spa rsity on the lattice. We then analyze neurophysiological data from high-level vi sual cortical area STSa, using an exact Bayesian approach to construct the forma l context needed by FCA. Prominent features of the resulting concept lattices ar e discussed, including indications for a product-of-experts code in real neurons .

****

Global Ranking Using Continuous Conditional Random Fields
Tao Qin, Tie-yan Liu, Xu-dong Zhang, De-sheng Wang, Hang Li
This paper studies global ranking problem by learning to rank methods. Conventio nal learning to rank methods are usually designed for `local ranking', in the se nse that the ranking model is defined on a single object, for example, a documen t in information retrieval. For many applications, this is a very loose approxim ation. Relations always exist between objects and it is better to define the ran king model as a function on all the objects to be ranked (i.e., the relations ar e also included). This paper refers to the problem as global ranking and propose s employing a Continuous Conditional Random Fields (CRF) for conducting the lear ning task. The Continuous CRF model is defined as a conditional probability dist ribution over ranking scores of objects conditioned on the objects. It can natur ally represent the content information of objects as well as the relation inform ation between objects, necessary for global ranking. Taking two specific informa tion retrieval tasks as examples, the paper shows how the Continuous CRF method can perform global ranking better than baselines.

****

Improving on Expectation Propagation
Manfred Opper, Ulrich Paquet, Ole Winther
We develop as series of corrections to Expectation Propagation (EP), which is on e of the most popular methods for approximate probabilistic inference. These cor rections can lead to improvements of the inference approximation or serve as a s anity check, indicating when EP yields unrealiable results.

****

Learning a discriminative hidden part model for human action recognition
Yang Wang, Greg Mori
We present a discriminative part-based approach for human action recognition fro m video sequences using motion features. Our model is based on the recently prop osed hidden conditional random field~(hCRF) for object recognition. Similar to h CRF for object recognition, we model a human action by a flexible constellation of parts conditioned on image observations. Different from object recognition, o ur model combines both large-scale global features and local patch features to d istinguish various actions. Our experimental results show that our model is comp arable to other state-of-the-art approaches in action recognition. In particular , our experimental results demonstrate that combining large-scale global feature s and local patch features performs significantly better than directly applying hCRF on local patches alone.

****

Adaptive Template Matching with Shift-Invariant Semi-NMF
Jonathan Roux, Alain Cheveigné, Lucas Parra
How does one extract unknown but stereotypical events that are linearly superimp osed within a signal with variable latencies and variable amplitudes? One could think of using template matching or matching pursuit to find the arbitrarily shi fted linear components. However, traditional matching approaches require that th e templates be known a priori. To overcome this restriction we use instead semi Non-Negative Matrix Factorization (semi-NMF) that we extend to allow for time sh ifts when matching the templates to the signal. The algorithm estimates template s directly from the data along with their non-negative amplitudes. The resulting method can be thought of as an adaptive template matching procedure. We demonst rate the procedure on the task of extracting spikes from single channel extracel

lular recordings. On these data the algorithm essentially performs spike detecti
on and unsupervised spike clustering. Results on simulated data and extracellula
r recordings indicate that the method performs well for signal-to-noise ratios o
f 6dB or higher and that spike templates are recovered accurately provided they
are sufficiently different.
****

Extracting State Transition Dynamics from Multiple Spike Trains with Correlated
Poisson HMM
Kentaro Katahira, Jun Nishikawa, Kazuo Okanoya, Masato Okada
Neural activity is non-stationary and varies across time. Hidden Markov Models (
HMMs) have been used to track the state transition among quasi-stationary discre
te neural states. Within this context, independent Poisson models have been used
 for the output distribution of HMMs; hence, the model is incapable of tracking
the change in correlation without modulating the firing rate. To achieve this, w
e applied a multivariate Poisson distribution with correlation terms for the out
put distribution of HMMs. We formulated a Variational Bayes (VB) inference for t
he model. The VB could automatically determine the appropriate number of hidden
states and correlation types while avoiding the overlearning problem. We develop
ed an efficient algorithm for computing posteriors using the recursive relations
hip of a multivariate Poisson distribution. We demonstrated the performance of o
ur method on synthetic data and a real spike train recorded from a songbird.
****

Partially Observed Maximum Entropy Discrimination Markov Networks
Jun Zhu, Eric Xing, Bo Zhang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Playing Pinball with non-invasive BCI
Matthias Krauledat, Konrad Grzeska, Max Sagebaum, Benjamin Blankertz, Carmen Vid
aurre, Klaus-Robert Müller, Michael Schröder
Compared to invasive Brain-Computer Interfaces (BCI), non-invasive BCI systems b
ased on Electroencephalogram (EEG) signals have not been applied successfully fo
r complex control tasks. In the present study, however, we demonstrate this is p
ossible and report on the interaction of a human subject with a complex real dev
ice: a pinball machine. First results in this single subject study clearly show
that fast and well-timed control well beyond chance level is possible, even thou
gh the environment is extremely rich and requires complex predictive behavior. U
sing machine learning methods for mental state decoding, BCI-based pinball contr
ol is possible within the first session without the necessity to employ lengthy
subject training. While the current study is still of anecdotal nature, it clear
ly shows that very compelling control with excellent timing and dynamics is poss
ible for a non-invasive BCI.
****

Logistic Normal Priors for Unsupervised Probabilistic Grammar Induction
Shay Cohen, Kevin Gimpel, Noah A. Smith
We explore a new Bayesian model for probabilistic grammars, a family of distribu
tions over discrete structures that includes hidden Markov models and probabilis
tic context-free grammars. Our model extends the correlated topic model framewor
k to probabilistic grammars, exploiting the logistic normal distribution as a pr
ior over the grammar parameters. We derive a variational EM algorithm for that m
odel, and then experiment with the task of unsupervised grammar induction for na
tural language dependency parsing. We show that our model achieves superior resu
lts over previous models that use different priors.
****

Risk Bounds for Randomized Sample Compressed Classifiers
Mohak Shah
We derive risk bounds for the randomized classifiers in Sample Compressions sett
ings where the classifier-specification utilizes two sources of information viz.

the compression set and the message string. By extending the recently proposed Occamâ■■s Hammer principle to the data-dependent settings, we derive point-wise versions of the bounds on the stochastic sample compressed classifiers and also recover the corresponding classical PAC-Bayes bound. We further show how these compare favorably to the existing results.

****

Learning the Semantic Correlation: An Alternative Way to Gain from Unlabeled Text

Yi Zhang, Artur Dubrawski, Jeff Schneider

In this paper, we address the question of what kind of knowledge is generally transferable from unlabeled text. We suggest and analyze the semantic correlation of words as a generally transferable structure of the language and propose a new method to learn this structure using an appropriately chosen latent variable model. This semantic correlation contains structural information of the language space and can be used to control the joint shrinkage of model parameters for any specific task in the same space through regularization. In an empirical study, we construct 190 different text classification tasks from a real-world benchmark, and the unlabeled documents are a mixture from all these tasks. We test the ability of various algorithms to use the mixed unlabeled text to enhance all classification tasks. Empirical results show that the proposed approach is a reliable and scalable method for semi-supervised learning, regardless of the source of unlabeled data, the specific task to be enhanced, and the prediction model used.

****

Online Optimization in X-Armed Bandits

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, Rémi Munos

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

****

Variational Mixture of Gaussian Process Experts

Chao Yuan, Claus Neubauer

Mixture of Gaussian processes models extended a single Gaussian process with ability of modeling multi-modal data and reduction of training complexity. Previous inference algorithms for these models are mostly based on Gibbs sampling, which can be very slow, particularly for large-scale data sets. We present a new generative mixture of experts model. Each expert is still a Gaussian process but is reformulated by a linear model. This breaks the dependency among training outputs and enables us to use a much faster variational Bayesian algorithm for training. Our gating network is more flexible than previous generative approaches as inputs for each expert are modeled by a Gaussian mixture model. The number of experts and number of Gaussian components for an expert are inferred automatically. A variety of tests show the advantages of our method.

****

On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost

Hamed Masnadi-shirazi, Nuno Vasconcelos

The machine learning problem of classifier design is studied from the perspective of probability elicitation, in statistics. This shows that the standard approach of proceeding from the specification of a loss, to the minimization of conditional risk is overly restrictive. It is shown that a better alternative is to start from the specification of a functional form for the minimum conditional risk, and derive the loss function. This has various consequences of practical interest, such as showing that 1) the widely adopted practice of relying on convex loss functions is unnecessary, and 2) many new losses can be derived for classification problems. These points are illustrated by the derivation of a new loss which is not convex, but does not compromise the computational tractability of classifier design, and is robust to the contamination of data with outliers. A new boosting algorithm, SavageBoost, is derived for the minimization of this loss. Experimental results show that it is indeed less sensitive to outliers than conven

tional methods, such as Ada, Real, or LogitBoost, and converges in fewer iterati
ons.
****

Non-stationary dynamic Bayesian networks
Joshua Robinson, Alexander Hartemink
A principled mechanism for identifying conditional dependencies in time-series d
ata is provided through structure learning of dynamic Bayesian networks (DBNs).
An important assumption of DBN structure learning is that the data are generated
 by a stationary processâ■■an assumption that is not true in many important sett
ings. In this paper, we introduce a new class of graphical models called non-sta
tionary dynamic Bayesian networks, in which the conditional dependence structure
 of the underlying data-generation process is permitted to change over time. Non
-stationary dynamic Bayesian networks represent a new framework for studying pro
blems in which the structure of a network is evolving over time. We define the n
on-stationary DBN model, present an MCMC sampling algorithm for learning the str
ucture of the model from time-series data under different assumptions, and demon
strate the effectiveness of the algorithm on both simulated and biological data.
****

Nonlinear causal discovery with additive noise models
Patrik Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, Bernhard Schölkopf
The discovery of causal relationships between a set of observed variables is a f
undamental problem in science. For continuous-valued data linear acyclic causal
models are often used because these models are well understood and there are wel
l-known methods to fit them to data. In reality, of course, many causal relation
ships are more or less nonlinear, raising some doubts as to the applicability an
d usefulness of purely linear methods. In this contribution we show that in fact
 the basic linear framework can be generalized to nonlinear models with additive
 noise. In this extended framework, nonlinearities in the data-generating proces
s are in fact a blessing rather than a curse, as they typically provide informat
ion on the underlying causal system and allow more aspects of the true data-gene
rating mechanisms to be identified. In addition to theoretical results we show s
imulations and some simple real data experiments illustrating the identification
 power provided by nonlinearities.
****

Simple Local Models for Complex Dynamical Systems
Erik Talvitie, Satinder Singh
We present a novel mathematical formalism for the idea of a local model,'' a mod
el of a potentially complex dynamical system that makes only certain predictions
 in only certain situations. As a result of its restricted responsibilities, a l
ocal model may be far simpler than a complete model of the system. We then show
how one might combine several local models to produce a more detailed model. We
demonstrate our ability to learn a collection of local models on a large-scale e
xample and do a preliminary empirical comparison of learning a collection of loc
al models and some other model learning methods."
****

Fitted Q-iteration by Advantage Weighted Regression
Gerhard Neumann, Jan Peters
Recently, fitted Q-iteration (FQI) based methods have become more popular due to
 their increased sample efficiency, a more stable learning process and the highe
r quality of the resulting policy. However, these methods remain hard to use for
 continuous action spaces which frequently occur in real-world tasks, e.g., in r
obotics and other technical applications. The greedy action selection commonly u
sed for the policy improvement step is particularly problematic as it is expensi
ve for continuous actions, can cause an unstable learning process, introduces an
 optimization bias and results in highly non-smooth policies unsuitable for real
-world systems. In this paper, we show that by using a soft-greedy action select
ion the policy improvement step used in FQI can be simplified to an inexpensive
advantage-weighted regression. With this result, we are able to derive a new, co
mputationally efficient FQI algorithm which can even deal with high dimensional
action spaces.

****
On the Generalization Ability of Online Strongly Convex Programming Algorithms
Sham M. Kakade, Ambuj Tewari
This paper examines the generalization properties of online convex programming a lgorithms when the loss function is Lipschitz and strongly convex. Our main resu lt is a sharp bound, that holds with high probability, on the excess risk of the output of an online algorithm in terms of the average regret. This allows one t o use recent algorithms with logarithmic cumulative regret guarantees to achieve fast convergence rates for the excess risk with high probability. The bound als o solves an open problem regarding the convergence rate of {\pegasos}, a recentl y proposed method for solving the SVM optimization problem.
****
Multiscale Random Fields with Application to Contour Grouping
Longin Latecki, Chengen Lu, Marc Sobel, Xiang Bai
We introduce a new interpretation of multiscale random fields (MSRFs) that admit s efficient optimization in the framework of regular (single level) random field s (RFs). It is based on a new operator, called append, that combines sets of ran dom variables (RVs) to single RVs. We assume that a MSRF can be decomposed into disjoint trees that link RVs at different pyramid levels. The append operator is then applied to map RVs in each tree structure to a single RV. We demonstrate t he usefulness of the proposed approach on a challenging task involving grouping contours of target shapes in images. MSRFs provide a natural representation of m ultiscale contour models, which are needed in order to cope with unstable contou r decompositions. The append operator allows us to find optimal image labels usi ng the classical framework of relaxation labeling, Alternative methods like Mark ov Chain Monte Carlo (MCMC) could also be used.
****
Modeling Short-term Noise Dependence of Spike Counts in Macaque Prefrontal Corte x
Arno Onken, Steffen Grünewälder, Matthias Munk, Klaus Obermayer
Correlations between spike counts are often used to analyze neural coding. The n oise is typically assumed to be Gaussian. Yet, this assumption is often inapprop riate, especially for low spike counts. In this study, we present copulas as an alternative approach. With copulas it is possible to use arbitrary marginal dist ributions such as Poisson or negative binomial that are better suited for modeli ng noise distributions of spike counts. Furthermore, copulas place a wide range of dependence structures at the disposal and can be used to analyze higher order interactions. We develop a framework to analyze spike count data by means of co pulas. Methods for parameter inference based on maximum likelihood estimates and for computation of Shannon entropy are provided. We apply the method to our dat a recorded from macaque prefrontal cortex. The data analysis leads to three sign ificant findings: (1) copula-based distributions provide better fits than discre tized multivariate normal distributions; (2) negative binomial margins fit the d ata better than Poisson margins; and (3) a dependence model that includes only p airwise interactions overestimates the information entropy by at least 19% compa red to the model with higher order interactions.
****
Predictive Indexing for Fast Search
Sharad Goel, John Langford, Alexander Strehl
We tackle the computational problem of query-conditioned search. Given a machine -learned scoring rule and a query distribution, we build a predictive index by p recomputing lists of potential results sorted based on an expected score of the result over future queries. The predictive index datastructure supports an anyti me algorithm for approximate retrieval of the top elements. The general approach is applicable to webpage ranking, internet advertisement, and approximate neare st neighbor search. It is particularly effective in settings where standard tech niques (e.g., inverted indices) are intractable. We experimentally find substant ial improvement over existing methods for internet advertisement and approximate nearest neighbors.
****

Human Active Learning
Rui Castro, Charles Kalish, Robert Nowak, Ruichen Qian, Tim Rogers, Jerry Zhu
We investigate a topic at the interface of machine learning and cognitive scienc e. Human active learning, where learners can actively query the world for inform ation, is contrasted with passive learning from random examples. Furthermore, we compare human active learning performance with predictions from statistical lea rning theory. We conduct a series of human category learning experiments inspire d by a machine learning task for which active and passive learning error bounds are well understood, and dramatically distinct. Our results indicate that humans are capable of actively selecting informative queries, and in doing so learn be tter and faster than if they are given random training data, as predicted by lea rning theory. However, the improvement over passive learning is not as dramatic as that achieved by machine active learning algorithms. To the best of our knowl edge, this is the first quantitative study comparing human category learning in active versus passive settings.
****

Clustered Multi-Task Learning: A Convex Formulation
Laurent Jacob, Jean-philippe Vert, Francis Bach
In multi-task learning several related tasks are considered simultaneously, with the hope that by an appropriate sharing of information across tasks, each task may benefit from the others. In the context of learning linear functions for sup ervised classification or regression, this can be achieved by including a priori information about the weight vectors associated with the tasks, and how they ar e expected to be related to each other. In this paper, we assume that tasks are clustered into groups, which are unknown beforehand, and that tasks within a gro up have similar weight vectors. We design a new spectral norm that encodes this a priori assumption, without the prior knowledge of the partition of tasks into groups, resulting in a new convex optimization formulation for multi-task learni ng. We show in simulations on synthetic examples and on the iedb MHC-I binding d ataset, that our approach outperforms well-known convex methods for multi-task l earning, as well as related non convex methods dedicated to the same problem.
****

Temporal Difference Based Actor Critic Learning - Convergence and Neural Impleme ntation
Dotan Castro, Dmitry Volkinshtein, Ron Meir
Actor-critic algorithms for reinforcement learning are achieving renewed popular ity due to their good convergence properties in situations where other approache s often fail (e.g., when function approximation is involved). Interestingly, the re is growing evidence that actor-critic approaches based on phasic dopamine sig nals play a key role in biological learning through the cortical and basal gangl ia. We derive a temporal difference based actor critic learning algorithm, for w hich convergence can be proved without assuming separate time scales for the act or and the critic. The approach is demonstrated by applying it to networks of sp iking neurons. The established relation between phasic dopamine and the temporal difference signal lends support to the biological relevance of such algorithms.
****

One sketch for all: Theory and Application of Conditional Random Sampling
Ping Li, Kenneth Church, Trevor Hastie
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.
****

The Mondrian Process
Daniel M. Roy, Yee Teh
We describe a novel stochastic process that can be used to construct a multidime nsional generalization of the stick-breaking process and which is related to the classic stick breaking process described by Sethuraman1994 in one dimension. We describe how the process can be applied to relational data modeling using the d e Finetti representation for infinitely and partially exchangeable arrays.

****
Scalable Algorithms for String Kernels with Inexact Matching
Pavel Kuksa, Pai-hsi Huang, Vladimir Pavlovic
We present a new family of linear time algorithms based on sufficient statistics for string comparison with mismatches under the string kernels framework. Our algorithms improve theoretical complexity bounds of existing approaches while scaling well with respect to the sequence alphabet size, the number of allowed mismatches and the size of the dataset. In particular, on large alphabets with loose mismatch constraints our algorithms are several orders of magnitude faster than the existing algorithms for string comparison under the mismatch similarity measure. We evaluate our algorithms on synthetic data and real applications in music genre classification, protein remote homology detection and protein fold prediction. The scalability of the algorithms allows us to consider complex sequence transformations, modeled using longer string features and larger numbers of mismatches, leading to a state-of-the-art performance with significantly reduced running times.
****