Ordinal Regression by Extended Binary Classification
Ling Li, Hsuan-tien Lin
We present a reduction framework from ordinal regression to binary classification based on extended examples. The framework consists of three steps: extracting extended examples from the original examples, learning a binary classifier on the extended examples with any binary classification algorithm, and constructing a ranking rule from the binary classifier. A weighted 0/1 loss of the binary classifier would then bound the mislabeling cost of the ranking rule. Our framework allows not only to design good ordinal regression algorithms based on well-tuned binary classification approaches, but also to derive new generalization bounds for ordinal regression from known bounds for binary classification. In addition, our framework unifies many existing ordinal regression algorithms, such as perceptron ranking and support vector ordinal regression. When compared empirically on benchmark data sets, some of our newly designed algorithms enjoy advantages in terms of both training speed and generalization performance over existing algorithms, which demonstrates the usefulness of our framework.
****
Learning to Traverse Image Manifolds
Piotr Dollár, Vincent Rabaud, Serge Belongie
We present a new algorithm, Locally Smooth Manifold Learning (LSML), that learns a warping function from a point on an manifold to its neighbors. Important characteristics of LSML include the ability to recover the structure of the manifold in sparsely populated regions and beyond the support of the provided data. Appli- cations of our proposed technique include embedding with a natural out-of-sample extension and tasks such as tangent distance estimation, frame rate up-conversion, video compression and motion transfer.
****
Game Theoretic Algorithms for Protein-DNA binding
Luis Pérez-breva, Luis E. Ortiz, Chen-hsiang Yeang, Tommi Jaakkola
We develop and analyze game-theoretic algorithms for predicting coordinate binding of multiple DNA binding regulators. The allocation of proteins to local neighborhoods and to sites is carried out with resource constraints while explicating competing and coordinate binding relations among proteins with affinity to the site or region. The focus of this paper is on mathematical foundations of the approach. We also briefly demonstrate the approach in the context of the -phage switch.
****
Kernel Maximum Entropy Data Transformation and an Enhanced Spectral Clustering Algorithm
Robert Jenssen, Torbjørn Eltoft, Mark Girolami, Deniz Erdogmus
We propose a new kernel-based data transformation technique. It is founded on the principle of maximum entropy (MaxEnt) preservation, hence named kernel MaxEnt. The key measure is Renyi's entropy estimated via Parzen windowing. We show that kernel MaxEnt is based on eigenvectors, and is in that sense similar to kernel PCA, but may produce strikingly different transformed data sets. An enhanced spectral clustering algorithm is proposed, by replacing kernel PCA by kernel MaxEnt as an intermediate step. This has a major impact on performance.
****
Unified Inference for Variational Bayesian Linear Gaussian State-Space Models
David Barber, Silvia Chiappa
Linear Gaussian State-Space Models are widely used and a Bayesian treatment of parameters is therefore of considerable interest. The approximate Variational Bayesian method applied to these models is an attractive approach, used successfully in applications ranging from acoustics to bioinformatics. The most challenging aspect of implementing the method is in performing inference on the hidden state sequence of the model. We show how to convert the inference problem so that standard Kalman Filtering/Smoothing recursions from the literature may be applied. This is in contrast to previously published approaches based on Belief Propagation. Our framework both simplifies and unifies the inference problem, so that future applications may be more easily developed. We demonstrate the elegance of t

he approach on Bayesian temporal ICA, with an application to finding independent dynamical processes underlying noisy EEG signals.

****

Graph Laplacian Regularization for Large-Scale Semidefinite Programming

Kilian Q. Weinberger, Fei Sha, Qihui Zhu, Lawrence Saul

In many areas of science and engineering, the problem arises how to discover low dimensional representations of high dimensional data. Recently, a number of researchers have converged on common solutions to this problem using methods from convex optimization. In particular, many results have been obtained by constructing semidefinite programs (SDPs) with low rank solutions. While the rank of matrix variables in SDPs cannot be directly constrained, it has been observed that low rank solutions emerge naturally by computing high variance or maximal trace solutions that respect local distance constraints. In this paper, we show how to solve very large problems of this type by a matrix factorization that leads to much smaller SDPs than those previously studied. The matrix factorization is derived by expanding the solution of the original problem in terms of the bottom eigenvectors of a graph Laplacian. The smaller SDPs obtained from this matrix factorization yield very good approximations to solutions of the original problem. Moreover, these approximations can be further refined by conjugate gradient descent. We illustrate the approach on localization in large scale sensor networks, where optimizations involving tens of thousands of nodes can be solved in just a few minutes.

****

Multi-Task Feature Learning

Andreas Argyriou, Theodoros Evgeniou, Massimiliano Pontil

We present a method for learning a low-dimensional representation which is shared across a set of multiple related tasks. The method builds upon the well- known 1-norm regularization problem using a new regularizer which controls the number of learned features common for all the tasks. We show that this problem is equivalent to a convex optimization problem and develop an iterative algorithm for solving it. The algorithm has a simple interpretation: it alternately performs a supervised and an unsupervised step, where in the latter step we learn common- across-tasks representations and in the former step we learn task-specific functions using these representations. We report experiments on a simulated and a real data set which demonstrate that the proposed method dramatically improves the per- formance relative to learning each task independently. Our algorithm can also be used, as a special case, to simply select – not learn – a few common features across the tasks.

****

Predicting spike times from subthreshold dynamics of a neuron

Ryota Kobayashi, Shigeru Shinomoto

It has been established that a neuron reproduces highly precise spike response to identical fluctuating input currents. We wish to accurately predict the firing times of a given neuron for any input current. For this purpose we adopt a model that mimics the dynamics of the membrane potential, and then take a cue from its dynamics for predicting the spike occurrence for a novel input current. It is found that the prediction is significantly improved by observing the state space of the membrane potential and its time derivative(s) in advance of a possible spike, in comparison to simply thresholding an instantaneous value of the estimated potential.

****

Blind source separation for over-determined delayed mixtures

Lars Omlor, Martin Giese

Blind source separation, i.e. the extraction of unknown sources from a set of given signals, is relevant for many applications. A special case of this problem is dimension reduction, where the goal is to approximate a given set of signals by superpositions of a minimal number of sources. Since in this case the signals outnumber the sources the problem is over-determined. Most popular approaches for addressing this problem are based on purely linear mixing models. However, many applications like the modeling of acoustic signals, EMG signals, or movement t rajectories, require temporal shift-invariance of the extracted components. This case has only rarely been treated in the computational literature, and specifically for the case of dimension reduction almost no algorithms have been proposed. We present a new algorithm for the solution of this problem, which is based on a timefrequency transformation (Wigner-Ville distribution) of the generative model. We show that this algorithm outperforms classical source separation algorithms for linear mixtures, and also a related method for mixtures with delays. In addition, applying the new algorithm to trajectories of human gaits, we demonstrate that it is suitable for the extraction of spatio-temporal components that are easier to interpret than components extracted with other classical algorithms.

****

Multiple timescales and uncertainty in motor adaptation

Konrad Körding, Joshua Tenenbaum, Reza Shadmehr

Our motor system changes due to causes that span multiple timescales. For example, muscle response can change because of fatigue, a condition where the disturbance has a fast timescale or because of disease where the disturbance is much slower. Here we hypothesize that the nervous system adapts in a way that reflects the temporal properties of such potential disturbances. According to a Bayesian formulation of this idea, movement error results in a credit assignment problem: what timescale is responsible for this disturbance? The adaptation schedule influences the behavior of the optimal learner, changing estimates at different timescales as well as the uncertainty. A system that adapts in this way predicts many properties observed in saccadic gain adaptation. It well predicts the timecourses of motor adaptation in cases of partial sensory deprivation and reversals of the adaptation direction.

****

Learning from Multiple Sources

Koby Crammer, Michael Kearns, Jennifer Wortman

We consider the problem of learning accurate models from multiple sources of "nearby" data. Given distinct samples from multiple data sources and estimates of the dissimilarities between these sources, we provide a general theory of which samples should be used to learn models for each source. This theory is applicable in a broad decision-theoretic learning framework, and yields results for classification and regression generally, and for density estimation within the exponential family. A key component of our approach is the development of approximate triangle inequalities for expected loss, which may be of independent interest.

****

Modeling Human Motion Using Binary Latent Variables

Graham W. Taylor, Geoffrey E. Hinton, Sam Roweis

We propose a non-linear generative model for human motion data that uses an undirected model with binary latent variables and real-valued "visible" variables that represent joint angles. The latent and visible variables at each time step receive directed connections from the visible variables at the last few time-steps. Such an architecture makes on-line inference efficient and allows us to use a simple approximate learning procedure. After training, the model finds a single set of parameters that simultaneously capture several different kinds of motion. We demonstrate the power of our approach by synthesizing various motion sequences and by performing on-line filling in of data lost during motion capture. Website: http://www.cs.toronto.edu/gwtaylor/publications/nips2006mhmublv/

****

Bayesian Image Super-resolution, Continued

Lyndsey Pickup, David Capel, Stephen J. Roberts, Andrew Zisserman

This paper develops a multi-frame image super-resolution approach from a Bayesian view-point by marginalizing over the unknown registration parameters relating the set of input low-resolution views. In Tipping and Bishop’s Bayesian image super-resolution approach [16], the marginalization was over the super- resolution image, necessitating the use of an unfavorable image prior. By inte- grating over the registration parameters rather than the high-resolution image, our method allows for more realistic prior distributions, and also reduces the dimen- sion of the integral considerably, removing the main computational bottleneck of the other algorithm. In addition to the motion model used by Tipping and Bishop, illumination components are introduced into the generative model, allowing us to handle changes in lighting as well as motion. We show results on real and synthetic datasets to illustrate the efficacy of this approach.

****

Geometric entropy minimization (GEM) for anomaly detection and localization
Alfred Hero

We introduce a novel adaptive non-parametric anomaly detection approach, called GEM, that is based on the minimal covering properties of K-point entropic graphs when constructed on N training samples from a nominal probability distribution. Such graphs have the property that as N   their span recovers the entropy minimizing set that supports at least = K/N (100)% of the mass of the Lebesgue part of the distribution. When a test sample falls outside of the entropy minimizing set an anomaly can be declared at a statistical level of significance = 1 - . A method for implementing this non-parametric anomaly detector is proposed that approximates this minimum entropy set by the influence region of a K-point entropic graph built on the training data. By implementing an incremental leave-one-out k-nearest neighbor graph on resampled subsets of the training data GEM can efficiently detect outliers at a given level of significance and compute their empirical p-values. We illustrate GEM for several simulated and real data sets in high dimensional feature spaces.

****

Bayesian Ensemble Learning
Hugh Chipman, Edward George, Robert Mcculloch

We develop a Bayesian "sum-of-trees" model, named BART, where each tree is constrained by a prior to be a weak learner. Fitting and inference are accomplished via an iterative backfitting MCMC algorithm. This model is motivated by ensemble methods in general, and boosting algorithms in particular. Like boosting, each weak learner (i.e., each weak tree) contributes a small amount to the overall model. However, our procedure is defined by a statistical model: a prior and a likelihood, while boosting is defined by an algorithm. This model-based approach enables a full and accurate assessment of uncertainty in model predictions, while remaining highly competitive in terms of predictive accuracy.

****

The Robustness-Performance Tradeoff in Markov Decision Processes
Huan Xu, Shie Mannor

Computation of a satisfactory control policy for a Markov decision process when the parameters of the model are not exactly known is a problem encountered in many practical applications. The traditional robust approach is based on a worstcase analysis and may lead to an overly conservative policy. In this paper we consider the tradeoff between nominal performance and the worst case performance over all possible models. Based on parametric linear programming, we propose a method that computes the whole set of Pareto efficient policies in the performancerobustness plane when only the reward parameters are subject to uncertainty. In the more general case when the transition probabilities are also subject to error, we show that the strategy with the "optimal" tradeoff might be non-Markovian and hence is in general not tractable.

****

Combining causal and similarity-based reasoning
Charles Kemp, Patrick Shafto, Allison Berke, Joshua Tenenbaum

Everyday inductive reasoning draws on many kinds of knowledge, including knowledge about relationships between properties and knowledge about relationships between objects. Previous accounts of inductive reasoning generally focus on just one kind of knowledge: models of causal reasoning often focus on relationships between properties, and models of similarity-based reasoning often focus on similarity relationships between objects. We present a Bayesian model of inductive reasoning that incorporates both kinds of knowledge, and show that it accounts well for human inferences about the properties of biological species.

****

Detecting Humans via Their Pose

Alessandro Bissacco, Ming-Hsuan Yang, Stefano Soatto

We consider the problem of detecting humans and classifying their pose from a single image. Specifically, our goal is to devise a statistical model that simultaneously answers two questions: 1) is there a human in the image? and, if so, 2) what is a low-dimensional representation of her pose? We investigate models that can be learned in an unsupervised manner on unlabeled images of human poses, and provide information that can be used to match the pose of a new image to the ones present in the training set. Starting from a set of descriptors recently proposed for human detection, we apply the Latent Dirichlet Allocation framework to model the statistics of these features, and use the resulting model to answer the above questions. We show how our model can efficiently describe the space of images of humans with their pose, by providing an effective representation of poses for tasks such as classification and matching, while performing remarkably well in human/non human decision problems, thus enabling its use for human detection. We validate the model with extensive quantitative experiments and comparisons with other approaches on human detection and pose matching.

****

Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space

Kyung-ah Sohn, Eric Xing

We present a new statistical framework called hidden Markov Dirichlet process (HMDP) to jointly model the genetic recombinations among possibly infinite number of founders and the coalescence-with-mutation events in the resulting genealogies. The HMDP posits that a haplotype of genetic markers is generated by a sequence of recombination events that select an ancestor for each locus from an unbounded set of founders according to a 1st-order Markov transition process. Conjoining this process with a mutation model, our method accommodates both between-lineage recombination and within-lineage sequence variations, and leads to a compact and natural interpretation of the population structure and inheritance process underlying haplotype data. We have developed an efficient sampling algo rithm for HMDP based on a two-level nested Polya urn scheme. On both simulated and real SNP haplotype data, our method performs competitively or significantly better than extant methods in uncovering the recombination hotspots along chromosomal loci; and in addition it also infers the ancestral genetic patterns and offers a highly accurate map of ancestral compositions of modern populations.

****

Online Classification for Complex Problems Using Simultaneous Projections

Yonatan Amit, Shai Shalev-shwartz, Yoram Singer

We describe and analyze an algorithmic framework for online classification where each online trial consists of multiple prediction tasks that are tied together. We tackle the problem of updating the online hypothesis by defining a projection problem in which each prediction task corresponds to a single linear constraint. These constraints are tied together through a single slack parameter. We then in- troduce a general method for approximately solving the problem by projecting simultaneously and independently on each constraint which corresponds to a prediction sub-problem, and then averaging the individual solutions. We show that this approach constitutes a feasible, albeit not necessarily optimal, solution for the original projection problem. We derive concrete simultaneous projection schemes and analyze them in the mistake bound model. We demonstrate the power of the proposed algorithm in experiments with online multiclass text categorization. Our experiments indicate that a combination of class-dependent features with the simultaneous projection method outperforms previously studied algorithms.

****

Convex Repeated Games and Fenchel Duality

Shai Shalev-shwartz, Yoram Singer

We describe an algorithmic framework for an abstract game which we term a convex repeated game. We show that various online learning and boosting algorithms can be all derived as special cases of our algorithmic framework. This unified view explains the properties of existing algorithms and also enables us to derive several new interesting algorithms. Our algorithmic framework stems from a connect ion that we build between the notions of regret in game theory and weak duality in convex optimization.

****

Towards a general independent subspace analysis

Fabian Theis

The increasingly popular independent component analysis (ICA) may only be applied to data following the generative ICA model in order to guarantee algorithmindependent and theoretically valid results. Subspace ICA models generalize the assumption of component independence to independence between groups of components. They are attractive candidates for dimensionality reduction methods, however are currently limited by the assumption of equal group sizes or less general semi-parametric models. By introducing the concept of irreducible independent subspaces or components, we present a generalization to a parameter-free mixture model. Moreover, we relieve the condition of at-most-one-Gaussian by including previous results on non-Gaussian component analysis. After introducing this general model, we discuss joint block diagonalization with unknown block sizes, on which we base a simple extension of JADE to algorithmically perform the subspace analysis. Simulations confirm the feasibility of the algorithm.

****

In-Network PCA and Anomaly Detection

Ling Huang, XuanLong Nguyen, Minos Garofalakis, Michael Jordan, Anthony Joseph, Nina Taft

We consider the problem of network anomaly detection in large distributed systems. In this setting, Principal Component Analysis (PCA) has been proposed as a method for discover- ing anomalies by continuously tracking the projection of the data onto a residual subspace. This method was shown to work well empirically in highly aggregated networks, that is, those with a limited number of large nodes and at coarse time scales. This approach, how- ever, has scalability limitations. To overcome these limitations, we develop a PCA-based anomaly detector in which adaptive local data (cid:2)lters send to a coordinator just enough data to enable accurate global detection. Our method is based on a stochastic matrix perturba- tion analysis that characterizes the tradeoff between the accuracy of anomaly detection and the amount of data communicated over the network.

****

Robotic Grasping of Novel Objects

Ashutosh Saxena, Justin Driemeyer, Justin Kearns, Andrew Ng

We consider the problem of grasping novel objects, specifically ones that are being seen for the first time through vision. We present a learning algorithm that neither requires, nor tries to build, a 3-d model of the object. Instead it predicts, directly as a function of the images, a point at which to grasp the object. Our algorithm is trained via supervised learning, using synthetic images for the training set. We demonstrate on a robotic manipulation platform that this approach successfully grasps a wide variety of objects, such as wine glasses, duct tape, markers, a translucent box, jugs, knife-cutters, cellphones, keys, screwdrivers, staplers, toothbrushes, a thick coil of wire, a strangely shaped power horn, and others, none of which were seen in the training set.

****

Bayesian Detection of Infrequent Differences in Sets of Time Series with Shared Structure

Jennifer Listgarten, Radford Neal, Sam Roweis, Rachel Puckrin, Sean Cutler

We present a hierarchical Bayesian model for sets of related, but different, classes of time series data. Our model performs alignment simultaneously across all classes, while detecting and characterizing class-specific differences. During inference the model produces, for each class, a distribution over a canonical representation of the class. These class-specific canonical representations are automatically aligned to one another -- preserving common sub-structures, and highlighting differences. We apply our model to compare and contrast solenoid valve current data, and also, liquid-chromatography-ultraviolet-diode array data from a study of the plant Arabidopsis thaliana.

****

Training Conditional Random Fields for Maximum Labelwise Accuracy

Samuel Gross, Olga Russakovsky, Chuong B., Serafim Batzoglou

We consider the problem of training a conditional random field (CRF) to maximize per-label predictive accuracy on a training set, an approach motivated by the principle of empirical risk minimization. We give a gradient-based procedure for minimizing an arbitrarily accurate approximation of the empirical risk under a Hamming loss function. In experiments with both simulated and real data, our optimization procedure gives significantly better testing performance than several current approaches for CRF training, especially in situations of high label noise.

****

Modeling Dyadic Data with Binary Latent Factors

Edward Meeds, Zoubin Ghahramani, Radford Neal, Sam Roweis

We introduce binary matrix factorization, a novel model for unsupervised ma- trix decomposition. The decomposition is learned by fitting a non-parametric Bayesian probabilistic model with binary latent variables to a matrix of dyadic data. Unlike bi-clustering models, which assign each row or column to a single cluster based on a categorical hidden feature, our binary feature model reflects the prior belief that items and attributes can be associated with more than one latent cluster at a time. We provide simple learning and inference rules for this new model and show how to extend it to an infinite model in which the number of features is not a priori fixed but is allowed to grow with the size of the data.

****

Reducing Calibration Time For Brain-Computer Interfaces: A Clustering Approach

Matthias Krauledat, Michael Schröder, Benjamin Blankertz, Klaus-Robert Müller

Up to now even subjects that are experts in the use of machine learning based BCI systems still have to undergo a calibration session of about 20-30 min. From this data their (movement) intentions are so far infered. We now propose a new paradigm that allows to completely omit such calibration and instead transfer knowledge from prior sessions. To achieve this goal we first define normalized CSP features and distances in-between. Second, we derive prototypical features across sessions: (a) by clustering or (b) by feature concatenation methods. Finally, we construct a classifier based on these individualized prototypes and show that, indeed, classifiers can be successfully transferred to a new session for a number of subjects.

****

Theory and Dynamics of Perceptual Bistability

Paul R. Schrater, Rashmi Sundareswara

Perceptual Bistability refers to the phenomenon of spontaneously switching between two or more interpretations of an image under continuous viewing. Although switching behavior is increasingly well characterized, the origins remain elusive. We propose that perceptual switching naturally arises from the brain's search for best interpretations while performing Bayesian inference. In particular, we propose that the brain explores a posterior distribution over image interpretations at a rapid time scale via a sampling-like process and updates its interpretation when a sampled interpretation is better than the discounted value of its current interpretation. We formalize the theory, explicitly derive switching rate distributions and discuss qualitative properties of the theory including the effect of changes in the posterior distribution on switching rates. Finally, predictions of the theory are shown to be consistent with measured changes in human switching dynamics to Necker cube stimuli induced by context.

****

Aggregating Classification Accuracy across Time: Application to Single Trial EEG

Steven Lemm, Christin Schäfer, Gabriel Curio

We present a method for binary on-line classification of triggered but temporally blurred events that are embedded in noisy time series in the context of on-line discrimination between left and right imaginary hand-movement. In particular the goal of the binary classification problem is to obtain the decision, as fast and as reliably as possible from the recorded EEG single trials. To provide a probabilistic decision at every time-point t the presented method gathers informat ion from two distinct sequences of features across time. In order to incorporate decisions from prior time-points we suggest an appropriate weighting scheme, that emphasizes time instances, providing a higher discriminatory power between the instantaneous class distributions of each feature, where the discriminatory power is quantified in terms of the Bayes error of misclassification. The effectiveness of this procedure is verified by its successful application in the 3rd BCI competition. Disclosure of the data after the competition revealed this approach to be superior with single trial error rates as low as 10.7, 11.5 and 16.7% for the three different sub jects under study.

****

Blind Motion Deblurring Using Image Statistics

Anat Levin

We address the problem of blind motion deblurring from a single image, caused by a few moving objects. In such situations only part of the image may be blurred, and the scene consists of layers blurred in different degrees. Most of of existing blind deconvolution research concentrates at recovering a single blurring kernel for the entire image. However, in the case of different motions, the blur cannot be modeled with a single kernel, and trying to deconvolve the entire image with the same kernel will cause serious artifacts. Thus, the task of deblurring needs to involve segmentation of the image into regions with different blurs. Our approach relies on the observation that the statistics of derivative filters in images are significantly changed by blur. Assuming the blur results from a constant velocity motion, we can limit the search to one dimensional box filter blurs. This enables us to model the expected derivatives distributions as a function of the width of the blur kernel. Those distributions are surprisingly powerful in discriminating regions with different blurs. The approach produces convincing deconvolution results on real world images with rich texture.

****

Accelerated Variational Dirichlet Process Mixtures

Kenichi Kurihara, Max Welling, Nikos Vlassis

Dirichlet Process (DP) mixture models are promising candidates for clustering applications where the number of clusters is unknown a priori. Due to compu- tational considerations these models are unfortunately unsuitable for large scale data-mining applications. We propose a class of deterministic accelerated DP mixture models that can routinely handle millions of data-cases. The speedup is achieved by incorporating kd-trees into a variational Bayesian algorithm for DP mixtures in the stick-breaking representation, similar to that of Blei and Jordan (2005). Our algorithm differs in the use of kd-trees and in the way we handle truncation: we only assume that the variational distributions are fixed at their pri- ors after a certain level. Experiments show that speedups relative to the standard variational algorithm can be significant.

****

Online Clustering of Moving Hyperplanes

René Vidal

We propose a recursive algorithm for clustering trajectories lying in multiple moving hyperplanes. Starting from a given or random initial condition, we use normalized gradient descent to update the coefficients of a time varying polynomial whose degree is the number of hyperplanes and whose derivatives at a trajectory give an estimate of the vector normal to the hyperplane containing that trajectory. As time proceeds, the estimates of the hyperplane normals are shown to track their true values in a stable fashion. The segmentation of the trajectories is then obtained by clustering their associated normal vectors. The final result is a simple recursive algorithm for segmenting a variable number of moving hyperplanes. We test our algorithm on the segmentation of dynamic scenes containing rigid motions and dynamic textures, e.g., a bird floating on water. Our method not only segments the bird motion from the surrounding water motion, but also determines patterns of motion in the scene (e.g., periodic motion) directly from the temporal evolution of the estimated polynomial coefficients. Our experiments also show that our method can deal with appearing and disappearing motions in the scene.

****

Efficient sparse coding algorithms

Honglak Lee, Alexis Battle, Rajat Raina, Andrew Ng

Sparse coding provides a class of algorithms for finding succinct representations of stimuli; given only unlabeled input data, it discovers basis functions that capture higher-level features in the data. However, finding sparse codes remains a very difficult computational problem. In this paper, we present efficient sparse coding algorithms that are based on iteratively solving two convex optimization problems: an L1 -regularized least squares problem and an L2 -constrained least squares problem. We propose novel algorithms to solve both of these optimization problems. Our algorithms result in a significant speedup for sparse coding, allowing us to learn larger sparse codes than possible with previously described algorithms. We apply these algorithms to natural images and demonstrate that the inferred sparse codes exhibit end-stopping and non-classical receptive field surround suppression and, therefore, may provide a partial explanation for these two phenomena in V1 neurons.

****

Approximate Correspondences in High Dimensions

Kristen Grauman, Trevor Darrell

Pyramid intersection is an efficient method for computing an approximate partial matching between two sets of feature vectors. We introduce a novel pyramid em- bedding based on a hierarchy of non-uniformly shaped bins that takes advantage of the underlying structure of the feature space and remains accurate even for sets with high-dimensional feature vectors. The matching similarity is computed in linear time and forms a Mercer kernel. Whereas previous matching approxima- tion algorithms suffer from distortion factors that increase linearly with the feature dimension, we demonstrate that our approach can maintain constant accuracy even as the feature dimension increases. When used as a kernel in a discriminative classifier, our approach achieves improved object recognition results over a state-of-the-art set kernel.

****

Temporal Coding using the Response Properties of Spiking Neurons

Thomas Voegtlin

In biological neurons, the timing of a spike depends on the timing of synaptic currents, in a way that is classically described by the Phase Response Curve. This has implications for temporal coding: an action potential that arrives on a synapse has an implicit meaning, that depends on the position of the postsynaptic neuron on the firing cycle. Here we show that this implicit code can be used to perform computations. Using theta neurons, we derive a spike-timing dependent learning rule from an error criterion. We demonstrate how to train an a uto-encoder neural network using this rule.

****

A Nonparametric Bayesian Method for Inferring Features From Similarity Judgments

Daniel Navarro, Thomas Griffiths

The additive clustering model is widely used to infer the features of a set of stimuli from their similarities, on the assumption that similarity is a weighted linear function of common features. This paper develops a fully Bayesian formulation of the additive clustering model, using methods from nonparametric Bayesian statistics to allow the number of features to vary. We use this to explore several approaches to parameter estimation, showing that the nonparametric Bayesian approach provides a straightforward way to obtain estimates of both the number of features used in producing similarity judgments and their importance.

****

Hierarchical Dirichlet Processes with Random Effects

Seyoung Kim, Padhraic Smyth

Data sets involving multiple groups with shared characteristics frequently arise in practice. In this paper we extend hierarchical Dirichlet processes to model such data. Each group is assumed to be generated from a template mixture model with group level variability in both the mixing proportions and the component parameters. Variabilities in mixing proportions across groups are handled using hie rarchical Dirichlet processes, also allowing for automatic determination of the number of components. In addition, each group is allowed to have its own component parameters coming from a prior described by a template mixture model. This group-level variability in the component parameters is handled using a random effects model. We present a Markov Chain Monte Carlo (MCMC) sampling algo- rithm to estimate model parameters and demonstrate the method by applying it to the problem of modeling spatial brain activation patterns across multiple images collected via functional magnetic resonance imaging (fMRI).

****

Bayesian Model Scoring in Markov Random Fields

Sridevi Parise, Max Welling

Scoring structures of undirected graphical models by means of evaluating the marginal likelihood is very hard. The main reason is the presence of the parti- tion function which is intractable to evaluate, let alone integrate over. We propose to approximate the marginal likelihood by employing two levels of approximation: we assume normality of the posterior (the Laplace approximation) and approximate all remaining intractable quantities using belief propagation and the linear response approximation. This results in a fast procedure for model scoring. Em- pirically, we find that our procedure has about two orders of magnitude better accuracy than standard BIC methods for small datasets, but deteriorates when the size of the dataset grows.

****

Isotonic Conditional Random Fields and Local Sentiment Flow

Yi Mao, Guy Lebanon

We examine the problem of predicting local sentiment flow in documents, and its application to several areas of text analysis. Formally, the problem is stated as predicting an ordinal sequence based on a sequence of word sets. In the spirit of isotonic regression, we develop a variant of conditional random fields that is well suited to handle this problem. Using the Mobius transform, we express the model as a simple convex optimization problem. Experiments demonstrate the model and its applications to sentiment prediction, style analysis, and text summarization.

****

Logistic Regression for Single Trial EEG Classification

Ryota Tomioka, Kazuyuki Aihara, Klaus-Robert Müller

We propose a novel framework for the classification of single trial ElectroEncephaloGraphy (EEG), based on regularized logistic regression. Framed in this robust statistical framework no prior feature extraction or outlier removal is required. We present two variations of parameterizing the regression function: (a) with a full rank symmetric matrix coefficient and (b) as a difference of two rank=1 matrices. In the first case, the problem is convex and the logistic regression is optimal under a generative model. The latter case is shown to be related to the Common Spatial Pattern (CSP) algorithm, which is a popular technique in Brain Computer Interfacing. The regression coefficients can also be topographically mapped onto the scalp similarly to CSP pro jections, which allows neuro-physiological interpretation. Simulations on 162 BCI datasets demonstrate that classification accuracy and robustness compares favorably against conventional CSP based classifiers.

****

A Local Learning Approach for Clustering

Mingrui Wu, Bernhard Schölkopf

We present a local learning approach for clustering. The basic idea is that a good clustering result should have the property that the cluster label of each data point can be well predicted based on its neighboring data and their cluster labels, using current supervised learning methods. An optimization problem is formulated such that its solution has the above property. Relaxation and eigen-decomposition are applied to solve this optimization problem. We also briefly investigate the parameter selection issue and provide a simple parameter selection method for the proposed algorithm. Experimental results are provided to validate the effectiveness of the proposed approach.

****

Temporal and Cross-Subject Probabilistic Models for fMRI Prediction Tasks

Alexis Battle, Gal Chechik, Daphne Koller

We present a probabilistic model applied to the fMRI video rating prediction task of the Pittsburgh Brain Activity Interpretation Competition (PBAIC) [2]. Our goal is to predict a time series of subjective, semantic ratings of a movie given functional MRI data acquired during viewing by three subjects. Our method uses conditionally trained Gaussian Markov random fields, which model both the relationships between the subjects' fMRI voxel measurements and the ratings, as well as the dependencies of the ratings across time steps and between subjects. We also employed non-traditional methods for feature selection and regularization that exploit the spatial structure of voxel activity in the brain. The model displayed good performance in predicting the scored ratings for the three subjects in test data sets, and a variant of this model was the third place entrant to the 2006 PBAIC.

****

Particle Filtering for Nonparametric Bayesian Matrix Factorization

Frank Wood, Thomas Griffiths

Many unsupervised learning problems can be expressed as a form of matrix factorization, reconstructing an observed data matrix as the product of two matrices of latent variables. A standard challenge in solving these problems is determining the dimensionality of the latent matrices. Nonparametric Bayesian matrix factorization is one way of dealing with this challenge, yielding a posterior distribution over possible factorizations of unbounded dimensionality. A drawback to this approach is that posterior estimation is typically done using Gibbs sampling, which can be slow for large problems and when conjugate priors cannot be used. As an alternative, we present a particle filter for posterior estimation in nonparametric Bayesian matrix factorization models. We illustrate this approach with two matrix factorization models and show favorable performance relative to Gibbs sampling.

****

Large-Scale Sparsified Manifold Regularization

Ivor Tsang, James Kwok

Semi-supervised learning is more powerful than supervised learning by using both labeled and unlabeled data. In particular, the manifold regularization framework, together with kernel methods, leads to the Laplacian SVM (LapSVM) that has demonstrated state-of-the-art performance. However, the LapSVM solution typically involves kernel expansions of all the labeled and unlabeled examples, and is slow on testing. Moreover, existing semi-supervised learning methods, including the LapSVM, can only handle a small number of unlabeled examples. In this paper, we integrate manifold regularization with the core vector machine, which has been used for large-scale supervised and unsupervised learning. By using a sparsified manifold regularizer and formulating as a center-constrained minimum enclosing ball problem, the proposed method produces sparse solutions with low time and space complexities. Experimental results show that it is much faster than the LapSVM, and can handle a million unlabeled examples on a standard PC; while the LapSVM can only handle several thousand patterns.

****

Non-rigid point set registration: Coherent Point Drift

Andriy Myronenko, Xubo Song, Miguel Carreira-Perpiñán

We introduce Coherent Point Drift (CPD), a novel probabilistic method for nonrigid registration of point sets. The registration is treated as a Maximum Likelihood (ML) estimation problem with motion coherence constraint over the velocity field such that one point set moves coherently to align with the second set. We formulate the motion coherence constraint and derive a solution of regularized ML estimation through the variational approach, which leads to an elegant kernel form. We also derive the EM algorithm for the penalized ML optimization with deterministic annealing. The CPD method simultaneously finds both the non-rigid transformation and the correspondence between two point sets without making any prior assumption of the transformation model except that of motion coherence. This me thod can estimate complex non-linear non-rigid transformations, and is shown to be accurate on 2D and 3D examples and robust in the presence of outliers and missing points.

****

Learning Nonparametric Models for Probabilistic Imitation
David Grimes, Daniel Rashid, Rajesh PN Rao

Learning by imitation represents an important mechanism for rapid acquisition of new behaviors in humans and robots. A critical requirement for learning by imitation is the ability to handle uncertainty arising from the observation process as well as the imitator's own dynamics and interactions with the environment. In this paper, we present a new probabilistic method for inferring imitative actions that takes into account both the observations of the teacher as well as the imitator's dynamics. Our key contribution is a nonparametric learning method which generalizes to systems with very different dynamics. Rather than relying on a known forward model of the dynamics, our approach learns a nonparametric forward model via exploration. Leveraging advances in approximate inference in graphical models, we show how the learned forward model can be directly used to plan an imitating sequence. We provide experimental results for two systems: a biomechanical model of the human arm and a 25-degrees-of-freedom humanoid robot. We demonstrate that the proposed method can be used to learn appropriate motor inputs to the model arm which imitates the desired movements. A second set of results demonstrates dynamically stable full-body imitation of a human teacher by the humanoid robot.

****

Active learning for misspecified generalized linear models
Francis Bach

Active learning refers to algorithmic frameworks aimed at selecting training data points in order to reduce the number of required training data points and/or im- prove the generalization performance of a learning method. In this paper, we present an asymptotic analysis of active learning for generalized linear models. Our analysis holds under the common practical situation of model misspecifica- tion, and is based on realistic assumptions regarding the nature of the sampling distributions, which are usually neither independent nor identical. We derive un- biased estimators of generalization performance, as well as estimators of expected reduction in generalization error after adding a new training data point, that allow us to optimize its sampling distribution through a convex optimization problem. Our analysis naturally leads to an algorithm for sequential active learning which is applicable for all tasks supported by generalized linear models (e.g., binary clas- sification, multi-class classification, regression) and can be applied in non-linear settings through the use of Mercer kernels.

****

Tighter PAC-Bayes Bounds
Amiran Ambroladze, Emilio Parrado-hernández, John Shawe-taylor

This paper proposes a PAC-Bayes bound to measure the performance of Support Vector Machine (SVM) classifiers. The bound is based on learning a prior over the distribution of classifiers with a part of the training samples. Experimental work shows that this bound is tighter than the original PAC-Bayes, resulting in an enhancement of the predictive capabilities of the PAC-Bayes bound. In addition, it is shown that the use of this bound as a means to estimate the hyperparameters of the classifier compares favourably with cross validation in terms of accuracy of the model, while saving a lot of computational burden.

****

Statistical Modeling of Images with Fields of Gaussian Scale Mixtures
Siwei Lyu, Eero Simoncelli

The local statistical properties of photographic images, when represented in a multi-scale basis, have been described using Gaussian scale mixtures (GSMs). Here, we use this local description to construct a global field of Gaussian scale mixtures (FoGSM). Specifically, we model subbands of wavelet coefficients as a product of an exponentiated homogeneous Gaussian Markov random field (hGMRF) and a second independent hGMRF. We show that parameter estimation for FoGSM is feasibl e, and that samples drawn from an estimated FoGSM model have marginal and joint statistics similar to wavelet coefficients of photographic images. We develop an algorithm for image denoising based on the FoGSM model, and demonstrate substantial improvements over current state-ofthe-art denoising method based on the local GSM model. Many successful methods in image processing and computer vision rely on statistical models for images, and it is thus of continuing interest to develop improved models, both in terms of their ability to precisely capture image structures, and in terms of their tractability when used in applications. Constructing such a model is difficult, primarily because of the intrinsic high dimensionality of the space of images. Two simplifying assumptions are usually made to reduce model complexity. The first is Markovianity: the density of a pixel conditioned on a small neighborhood, is assumed to be independent from the rest of the image. The second assumption is homogeneity: the local density is assumed to be independent of its absolute position within the image. The set of models satisfying both of these assumptions constitute the class of homogeneous Markov random fields (hMRFs). Over the past two decades, studies of photographic images represented with multi-scale multiorientation image decompositions (loosely referred to as "wavelets") have revealed striking nonGaussian regularities and inter and intra-subband dependencies. For instance, wavelet coefficients generally have highly kurtotic marginal distributions [1, 2], and their amplitudes exhibit strong correlations with the amplitudes of nearby coefficients [3, 4]. One model that can capture the nonGaussian marginal behaviors is a product of non-Gaussian scalar variables [5]. A number of authors have developed non-Gaussian MRF models based on this sort of local description [6, 7, 8], among which the recently developed fields of experts model [7] has demonstrated impressive performance in denoising (albeit at an extremely high computational cost in learning model parameters). An alternative model that can capture non-Gaussian local structure is a scale mixture model [9, 10, 11]. An important special case is Gaussian scale mixtures (GSM), which consists of a Gaussian random vector whose amplitude is modulated by a hidden scaling variable. The GSM model provides a particularly good description of local image statistics, and the Gaussian substructure of the model leads to efficient algorithms for parameter estimation and inference. Local GSM-based methods represent the current state-of-the-art in image denoising [12]. The power of GSM models should be substantially improved when extended to describe more than a small neighborhood of wavelet coefficients. To this end, several authors have embedded local Gaussian mixtures into tree-structured

****

Near-Uniform Sampling of Combinatorial Spaces Using XOR Constraints

Carla P. Gomes, Ashish Sabharwal, Bart Selman

We propose a new technique for sampling the solutions of combinatorial problems in a near-uniform manner. We focus on problems specified as a Boolean formula, i.e., on SAT instances. Sampling for SAT problems has been shown to have interesting connections with probabilistic reasoning, making practical sampling algorithms for SAT highly desirable. The best current approaches are based on Markov Chain Monte Carlo methods, which have some practical limitations. Our approach exploits combinatorial properties of random parity (X O R) constraints to prune away solutions near-uniformly. The final sample is identified amongst the remaining ones using a state-of-the-art SAT solver. The resulting sampling distribution is provably arbitrarily close to uniform. Our experiments show that our technique achieves a significantly better sampling quality than the best alternative.

****

Recursive Attribute Factoring

David Cohn, Deepak Verma, Karl Pfleger

Clustering, or factoring of a document collection attempts to “explain” each observed document in terms of one or a small number of inferred prototypes. Prior work demonstrated that when links exist between documents in the corpus (as is the case with a collection of web pages or scientific papers), building a joint model of document contents and connections produces a better model than that built from contents or connections alone. Many problems arise when trying to apply these joint models to corpus at the scale of the World Wide Web, however; one of these is that the sheer overhead of representing a feature space on the order of billions of dimensions becomes impractical. We address this problem with a simple representational shift inspired by proba- bilistic relational models: instead of representing document linkage in terms of the identities of linking documents, we represent it by the explicit and inferred at- tributes of the linking documents. Several surprising results come with this shift: in addition to being computationally more tractable, the new model produces fac- tors that more cleanly decompose the document collection. We discuss several variations on this model and show how some can be seen as exact generalizations of the PageRank algorithm.

****

Information Bottleneck for Non Co-Occurrence Data

Yevgeny Seldin, Noam Slonim, Naftali Tishby

We present a general model-independent approach to the analysis of data in cases when these data do not appear in the form of co-occurrence of two variables X, Y , but rather as a sample of values of an unknown (stochastic) function Z (X, Y). For example, in gene expression data, the expression level Z is a function of gene X and condition Y ; or in movie ratings data the rating Z is a function of viewer X and movie Y . The approach represents a consistent extension of the Information Bottleneck method that has previously relied on the availability of co-occurrence statistics. By altering the relevance variable we eliminate the need in the sample of joint distribution of all input variables. This new formulation also enables simple MDL-like model complexity control and prediction of missing values of Z . The approach is analyzed and shown to be on a par with the best known clustering algorithms for a wide range of domains. For the prediction of missing values (collaborative filtering) it improves the currently best known results.

****

A Probabilistic Algorithm Integrating Source Localization and Noise Suppression of MEG and EEG data

Johanna Zumer, Hagai Attias, Kensuke Sekihara, Srikantan Nagarajan

We have developed a novel algorithm for integrating source localization and noise suppression based on a probabilistic graphical model of stimulus-evoked MEG/EEG data. Our algorithm localizes multiple dipoles while suppressing noise sources with the computational complexity equivalent to a single dipole scan, and is therefore more ef(cid:2)cient than traditional multidipole (cid:2)tting procedures. In simulation, the algorithm can accurately localize and estimate the time course of several simultaneously-active dipoles, with rotating or (cid:2)xed orientation, at noise levels typical for averaged MEG data. Furthermore, the algorithm is superior to beamforming techniques, which we show to be an approximation to our graphical model, in estimation of temporally correlated sources. Success of this algorithm for localizing auditory cortex in a tumor patient and for localizing an epileptic spike source are also demonstrated.

****

Attentional Processing on a Spike-Based VLSI Neural Network

Yingxue Wang, Rodney Douglas, Shih-Chii Liu

The neurons of the neocortex communicate by asynchronous events called action potentials (or 'spikes'). However, for simplicity of simulation, most models of processing by cortical neural networks have assumed that the activations of their neurons can be approximated by event rates rather than taking account of individual spikes. The obstacle to exploring the more detailed spike processing of these networks has been reduced considerably in recent years by the development of hybrid analog-digital Very-Large Scale Integrated (hVLSI) neural networks composed of spiking neurons that are able to operate in real-time. In this paper we describe such a hVLSI neural network that performs an interesting task of selective attentional processing that was previously described for a simulated 'pointer-map' rate model by Hahnloser and colleagues. We found that most of the computational features of their rate model can be reproduced in the spiking implementation; but, that spike-based processing requires a modification of the original network architecture in order to memorize a previously attended target.

****

A Bayesian Approach to Diffusion Models of Decision-Making and Response Time
Michael Lee, Ian Fuss, Daniel Navarro

We present a computational Bayesian approach for Wiener diffusion models, which are prominent accounts of response time distributions in decision-making. We first develop a general closed-form analytic approximation to the response time distributions for one-dimensional diffusion processes, and derive the required Wiener diffusion as a special case. We use this result to undertake Bayesian modeling of benchmark data, using posterior sampling to draw inferences about the interesting psychological parameters. With the aid of the benchmark data, we show the Bayesian account has several advantages, including dealing naturally with the parameter variation needed to account for some key features of the data, and providing quantitative measures to guide decisions about model construction.

****

Graph-Based Visual Saliency
Jonathan Harel, Christof Koch, Pietro Perona

A new bottom-up visual saliency model, Graph-Based Visual Saliency (GBVS), is proposed. It consists of two steps: first forming activation maps on certain feature channels, and then normalizing them in a way which highlights conspicuity and admits combination with other maps. The model is simple, and biologically plausible insofar as it is naturally parallelized. This model powerfully predicts human fixations on 749 variations of 108 natural images, achieving 98% of the ROC area of a human-based control, whereas the classical algorithms of Itti & Koch ([2], [3], [4]) achieve only 84%.

****

Doubly Stochastic Normalization for Spectral Clustering
Ron Zass, Amnon Shashua

In this paper we focus on the issue of normalization of the affinity matrix in spectral clustering. We show that the difference between N-cuts and Ratio-cuts is in the error measure being used (relative-entropy versus L1 norm) in finding the closest doubly-stochastic matrix to the input affinity matrix. We then develop a scheme for finding the optimal, under Frobenius norm, doubly-stochastic approximation using Von-Neumann's successive projections lemma. The new normalization scheme is simple and efficient and provides superior clustering performance over many of the standardized tests.

****

A Scalable Machine Learning Approach to Go
Lin Wu, Pierre Baldi

Go is an ancient board game that poses unique opportunities and challenges for AI and machine learning. Here we develop a machine learning approach to Go, and related board games, focusing primarily on the problem of learning a good eval- uation function in a scalable way. Scalability is essential at multiple levels, from the library of local tactical patterns, to the integration of patterns across the board, to the size of the board itself. The system we propose is capable of automatically learning the propensity of local patterns from a library of games. Propensity and other local tactical information are fed into a recursive neural network, derived from a Bayesian network architecture. The network integrates local information across the board and produces local outputs that represent local territory owner- ship probabilities. The aggregation of these probabilities provides an effective strategic evaluation function that is an estimate of the expected area at the end (or at other stages) of the game. Local area targets for training can be derived from datasets of human games. A system trained using only 9 × 9 amateur game data performs surprisingly well on a test set derived from 19 × 19 professional game data. Possible directions for further improvements are briefly discussed.

****

Stratification Learning: Detecting Mixed Density and Dimensionality in High Dimensional Point Clouds
Gloria Haro, Gregory Randall, Guillermo Sapiro

The study of point cloud data sampled from a stratification, a collection of manifolds with possible different dimensions, is pursued in this paper. We present a technique for simultaneously soft clustering and estimating the mixed dimensionality and density of such structures. The framework is based on a maximum likelihood estimation of a Poisson mixture model. The presentation of the approach is completed with artificial and real examples demonstrating the importance of extending manifold learning to stratification learning.

****

A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation

Yee Teh, David Newman, Max Welling

Latent Dirichlet allocation (LDA) is a Bayesian network that has recently gained much popularity in applications ranging from document modeling to computer vision. Due to the large scale nature of these applications, current inference procedures like variational Bayes and Gibbs sampling have been found lacking. In this paper we propose the collapsed variational Bayesian inference algorithm for LDA, and show that it is computationally efficient, easy to implement and significantly more accurate than standard variational Bayesian inference for LDA.

****

Dynamic Foreground/Background Extraction from Images and Videos using Random Patches

Le Lu, Gregory Hager

In this paper, we propose a novel exemplar-based approach to extract dynamic foreground regions from a changing background within a collection of images or a video sequence. By using image segmentation as a pre-processing step, we convert this traditional pixel-wise labeling problem into a lower-dimensional supervised, binary labeling procedure on image segments. Our approach consists of three steps. First, a set of random image patches are spatially and adaptively sampled within each segment. Second, these sets of extracted samples are formed into two "bags of patches" to model the foreground/background appearance, respectively. We perform a novel bidirectional consistency check between new patches from incoming frames and current "bags of patches" to reject outliers, control model rigidity and make the model adaptive to new observations. Within each bag, image patches are further partitioned and resampled to create an evolving appearance model. Finally, the foreground/background decision over segments in an image is formulated using an aggregation function defined on the similarity measurements of sampled patches relative to the foreground and background models. The essence of the algorithm is conceptually simple and can be easily implemented within a few hundred lines of Matlab code. We evaluate and validate the proposed approach by extensive real examples of the object-level image mapping and tracking within a variety of challenging environments. We also show that it is straightforward to apply our problem formulation on non-rigid object tracking with difficult surveillance videos.

****

Subordinate class recognition using relational object models

Aharon Hillel, Daphna Weinshall

We address the problem of sub-ordinate class recognition, like the distinction between different types of motorcycles. Our approach is motivated by observations from cognitive psychology, which identify parts as the defining component of basic level categories (like motorcycles), while sub-ordinate categories are more often defined by part properties (like 'jagged wheels'). Accordingly, we suggest a two-stage algorithm: First, a relational part based object model is learnt using unsegmented object images from the inclusive class (e.g., motorcycles in general). The model is then used to build a class-specific vector representation for images, where each entry corresponds to a model's part. In the second stage we train a standard discriminative classifier to classify subclass instances (e.g., cross motorcycles) based on the class-specific vector representation. We describe extensive experimental results with several subclasses. The proposed algorithm typically gives better results than a competing one-step algorithm, or a two stage algorithm where classification is based on a model of the sub-ordinate class.

****

Dirichlet-Enhanced Spam Filtering based on Biased Samples
Steffen Bickel, Tobias Scheffer

We study a setting that is motivated by the problem of filtering spam messages for many users. Each user receives messages according to an individual, unknown distribution, reflected only in the unlabeled inbox. The spam filter for a user is required to perform well with respect to this distribution. Labeled messages from publicly available sources can be utilized, but they are governed by a distinct distribution, not adequately representing most inboxes. We devise a method that minimizes a loss function with respect to a user's personal distribution based on the available biased sample. A nonparametric hierarchical Bayesian model furthermore generalizes across users by learning a common prior which is imposed on new email accounts. Empirically, we observe that bias-corrected learning outperforms naive reliance on the assumption of independent and identically distributed data; Dirichlet-enhanced generalization across users outperforms a single ("one size fits all") filter as well as independent filters for all users.

****

Stability of K-Means Clustering
Alexander Rakhlin, Andrea Caponnetto

We phrase K -means clustering as an empirical risk minimization procedure over a class HK and explicitly calculate the covering number for this class. Next, we show that stability of K -means clustering is characterized by the geometry of HK with respect to the underlying distribution. We prove that in the case of a unique global minimizer, the clustering solution is stable with respect to complete changes of the data, while for the case of multiple minimizers, the change of (n1/2) samples defines the transition between stability and instability. While for a finite number of minimizers this result follows from multinomial distribution estimates, the case of infinite minimizers requires more refined tools. We conclude by proving that stability of the functions in HK implies stability of the actual centers of the clusters. Since stability is often used for selecting the number of clusters in practice, we hope that our analysis serves as a starting point for finding theoretically grounded recipes for the choice of K .

****

Convergence of Laplacian Eigenmaps
Mikhail Belkin, Partha Niyogi

Geometrically based methods for various tasks of machine learning have attracted considerable attention over the last few years. In this paper we show convergence of eigenvectors of the point cloud Laplacian to the eigen- functions of the Laplace-Beltrami operator on the underlying manifold, thus establishing the first convergence results for a spectral dimensionality re- duction algorithm in the manifold setting.

****

Bayesian Policy Gradient Algorithms
Mohammad Ghavamzadeh, Yaakov Engel

Policy gradient methods are reinforcement learning algorithms that adapt a param- eterized policy by following a performance gradient estimate. Conventional pol- icy gradient methods use Monte-Carlo techniques to estimate this gradient. Since Monte Carlo methods tend to have high variance, a large number of samples is required, resulting in slow convergence. In this paper, we propose a Bayesian framework that models the policy gradient as a Gaussian process. This reduces the number of samples needed to obtain accurate gradient estimates. Moreover, estimates of the natural gradient as well as a measure of the uncertainty in the gradient estimates are provided at little extra cost.

****

Handling Advertisements of Unknown Quality in Search Advertising
Sandeep Pandey, Christopher Olston

We consider how a search engine should select advertisements to display with search results, in order to maximize its revenue. Under the standard "pay-per-click" arrangement, revenue depends on how well the displayed advertisements appeal to users. The main difficulty stems from new advertisements whose degree of appeal has yet to be determined. Often the only reliable way of determining appeal is exploration via display to users, which detracts from exploitation of other advertisements known to have high appeal. Budget constraints and finite advertisement lifetimes make it necessary to explore as well as exploit. In this paper we study the tradeoff between exploration and exploitation, modeling advertisement placement as a multi-armed bandit problem. We extend traditional bandit formulations to account for budget constraints that occur in search engine advertising markets, and derive theoretical bounds on the performance of a family of algorithms. We measure empirical performance via extensive experiments over real-world data.

****

Part-based Probabilistic Point Matching using Equivalence Constraints

Graham Mcneill, Sethu Vijayakumar

Correspondence algorithms typically struggle with shapes that display part-based variation. We present a probabilistic approach that matches shapes using independent part transformations, where the parts themselves are learnt during matching. Ideas from semi-supervised learning are used to bias the algorithm towards finding `perceptually valid' part structures. Shapes are represented by unlabeled point sets of arbitrary size and a background component is used to handle occlusion, local dissimilarity and clutter. Thus, unlike many shape matching techniques, our approach can be applied to shapes extracted from real images. Model parameters are estimated using an EM algorithm that alternates between finding a soft correspondence and computing the optimal part transformations using Procrustes analysis.

****

Greedy Layer-Wise Training of Deep Networks

Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle

Recent analyses (Bengio, Delalleau, & Le Roux, 2006; Bengio & Le Cun, 2007) of modern nonparametric machine learning algorithms that are kernel machines, such as Support Vector Machines (SVMs), graph-based manifold and semi-supervised learning algorithms suggest fundamental limitations of some learning algorithms. The problem is clear in kernel-based approaches when the kernel is "local" (e.g., the Gaussian kernel), i.e., K (x, y) converges to a constant when ||x - y || increases. These analyses point to the difficulty of learning "highly-varying functions", i.e., functions that have a large number of "variations" in the domain of interest, e.g., they would require a large number of pieces to be well represented by a piecewise-linear approximation. Since the number of pieces can be made to grow exponentially with the number of factors of variations in the input, this is connected with the well-known curse of dimensionality for classical non-parametric learning algorithms (for regression, classification and density estimation). If the shapes of all these pieces are unrelated, one needs enough examples for each piece in order to generalize properly. However, if these shapes are related and can be predicted from each other, "non-local" learning algorithms have the potential to generalize to pieces not covered by the training set. Such ability would seem necessary for learning in complex domains such as Artificial Intelligence tasks (e.g., related to vision, language, speech, robotics). Kernel machines (not only those with a local kernel) have a shallow architecture, i.e., only two levels of data-dependent computational elements. This is also true of feedforward neural networks with a single hidden layer (which can become SVMs when the number of hidden units becomes large (Bengio, Le Roux, Vincent, Delalleau, & Marcotte, 2006)). A serious problem with shallow architectures is that they can be very inefficient in terms of the number of computational units (e.g., bases, hidden units), and thus in terms of required examples (Bengio & Le Cun, 2007). One way to represent a highly-varying function compactly (with few parameters) is through the composition of many non-linearities, i.e., with a deep architecture. For example, the parity function with d inputs requires O(2d) examples and parameters to be represented by a Gaussian SVM (Bengio et al., 2006), O(d2) parameters for a one-hidden-layer neural network, O(d) parameters and units for a multi-layer network with O(log2 d) layers, and O(1) parameters with a recurrent neural network. More generally,

****

Optimal Change-Detection and Spiking Neurons

Angela J. Yu

Survival in a non-stationary, potentially adversarial environment requires animals to detect sensory changes rapidly yet accurately, two oft competing desiderata. Neurons subserving such detections are faced with the corresponding challenge to discern "real" changes in inputs as quickly as possible, while ignoring noisy fluctuations. Mathematically, this is an example of a change-detection problem that is actively researched in the controlled stochastic processes community. In this paper, we utilize sophisticated tools developed in that community to formalize an instantiation of the problem faced by the nervous system, and characterize the Bayes-optimal decision policy under certain assumptions. We will derive from this optimal strategy an information accumulation and decision process that remarkably resembles the dynamics of a leaky integrate-and-fire neuron. This correspondence suggests that neurons are optimized for tracking input changes, and sheds new light on the computational import of intracellular properties such as resting membrane potential, voltage-dependent conductance, and post-spike reset voltage. We also explore the influence that factors such as timing, uncertainty, neuromodulation, and reward should and do have on neuronal dynamics and sensitivity, as the optimal decision strategy depends critically on these factors.

****

Temporal dynamics of information content carried by neurons in the primary visual cortex

Danko Nikoli■, Stefan Haeusler, Wolf Singer, Wolfgang Maass

We use multi-electrode recordings from cat primary visual cortex and investigate
whether a simple linear classifier can extract information about the presented stim(cid:173)
uli. We find that information is extractable and that it even lasts for several hun(cid:173)
dred milliseconds after the stimulus has been removed. In a fast sequence of stim(cid:173)
ulus presentation, information about both new and old stimuli is present simul(cid:173)
taneously and nonlinear relations between these stimuli can be extracted. These
results suggest nonlinear properties of cortical representations. The important im(cid:173)
plications of these properties for the nonlinear brain theory are discussed.

****

A Switched Gaussian Process for Estimating Disparity and Segmentation in Binocular Stereo

Oliver Williams

This paper describes a Gaussian process framework for inferring pixel-wise disparity and bi-layer segmentation of a scene given a stereo pair of images. The Gaussian process covariance is parameterized by a foreground-backgroundocclusion segmentation label to model both smooth regions and discontinuities. As such, we call our model a switched Gaussian process. We propose a greedy incremental algorithm for adding observations from the data and assigning segmentation labels. Two observation schedules are proposed: the first treats scanlines as independent, the second uses an active learning criterion to select a sparse subset of points to measure. We show that this probabilistic framework has comparable performance to the state-of-the-art.

****

Automated Hierarchy Discovery for Planning in Partially Observable Environments

Laurent Charlin, Pascal Poupart, Romy Shioda

Planning in partially observable domains is a notoriously dif■cult problem. However, in many real-world scenarios, planning can be simpli■ed by decomposing the task into a hierarchy of smaller planning problems. Several approaches have been proposed to optimize a policy that decomposes according to a hierarchy speci■ed a priori. In this paper, we investigate the problem of automatically discover ing the hierarchy. More precisely, we frame the optimization of a hierarchical policy as a non-convex optimization problem that can be solved with general non-linear solvers, a mixed-integer non-linear approximation or a form of bounded hierar- chical policy iteration. By encoding the hierarchical structure as variables of the optimization problem, we can automatically discover a hierarchy. Our method is flexible enough to allow any parts of the hierarchy to be specified based on prior knowledge while letting the optimization discover the unknown parts. It can also discover hierarchical policies, including recursive policies, that are more compact (potentially infinitely fewer parameters) and often easier to understand given the decomposition induced by the hierarchy.

****

Adaptor Grammars: A Framework for Specifying Compositional Nonparametric Bayesian Models

Mark Johnson, Thomas Griffiths, Sharon Goldwater

This paper introduces adaptor grammars, a class of probabilistic models of language that generalize probabilistic context-free grammars (PCFGs). Adaptor grammars augment the probabilistic rules of PCFGs with “adaptors” that can in- duce dependencies among successive uses. With a particular choice of adaptor, based on the Pitman-Yor process, nonparametric Bayesian models of language using Dirichlet processes and hierarchical Dirichlet processes can be written as simple grammars. We present a general-purpose inference algorithm for adaptor grammars, making it easy to define and use such models, and illustrate how several existing nonparametric Bayesian models can be expressed within this framework.

****

On Transductive Regression

Corinna Cortes, Mehryar Mohri

In many modern large-scale learning applications, the amount of unlabeled data far exceeds that of labeled data. A common instance of this problem is the transductive setting where the unlabeled test points are known to the learning algorithm. This paper presents a study of regression problems in that setting. It presents explicit VC-dimension error bounds for transductive regression that hold for all bounded loss functions and coincide with the tight classification bounds of Vapnik when applied to classification. It also presents a new transductive regression algorithm inspired by our bound that admits a primal and kernelized closedform solution and deals efficiently with large amounts of unlabeled data. The algorithm exploits the position of unlabeled points to locally estimate their labels and then uses a global optimization to ensure robust predictions. Our study also includes the results of experiments with several publicly available regression data sets with up to 20,000 unlabeled examples. The comparison with other transductive regression algorithms shows that it performs well and that it can scale to large data sets.

****

Large Margin Multi-channel Analog-to-Digital Conversion with Applications to Neural Prosthesis

Amit Gore, Shantanu Chakrabartty

A key challenge in designing analog-to-digital converters for cortically implanted prosthesis is to sense and process high-dimensional neural signals recorded by the micro-electrode arrays. In this paper, we describe a novel architecture for analog-to-digital (A/D) conversion that combines conversion with spatial de-correlation within a single module. The architecture called multiple-input multiple-output (MIMO) is based on a min-max gradient descent optimization of a regularized linear cost function that naturally lends to an A/D formulation. Using an online formulation, the architecture can adapt to slow variations in cross-channel correlations, observed due to relative motion of the microelectrodes with respect to the signal sources. Experimental results with real recorded multi-channel neural data demonstrate the effectiveness of the proposed algorithm in alleviating cross-channel redundancy across electrodes and performing data-compression directly at the A/D converter.

****

Comparative Gene Prediction using Conditional Random Fields

Jade Vinson, David Decaprio, Matthew Pearson, Stacey Luoma, James Galagan

Computational gene prediction using generative models has reached a plateau, with several groups converging to a generalized hidden Markov model (GHMM) incorporating phylogenetic models of nucleotide sequence evolution. Further improvements in gene calling accuracy are likely to come through new methods that incorporate additional data, both comparative and species specific. Conditional Random Fields (CRFs), which directly model the conditional probability P (y |x) of a vector of hidden states conditioned on a set of observations, provide a unified framework for combining probabilistic and non-probabilistic information and have been shown to outperform HMMs on sequence labeling tasks in natural language processing. We describe the use of CRFs for comparative gene prediction. We implement a model that encapsulates both a phylogenetic-GHMM (our baseline comparative model) and additional non-probabilistic features. We tested our model on the genome sequence of the fungal human pathogen Cryptococcus neoformans. Our baseline comparative model displays accuracy comparable to the the best available gene prediction tool for this organism. Moreover, we show that discriminative training and the incorporation of non-probabilistic evidence significantly improve performance. Our software implementation, Conrad, is freely available with an open source license at http://www.broad.mit.edu/annotation/conrad/.

****

Learning annotated hierarchies from relational data

Daniel M. Roy, Charles Kemp, Vikash Mansinghka, Joshua Tenenbaum

The objects in many real-world domains can be organized into hierarchies, where each internal node picks out a category of objects. Given a collection of fea- tures and relations defined over a set of objects, an annotated hierarchy includes a specification of the categories that are most useful for describing each individual feature and relation. We define a generative model for annotated hierarchies and the features and relations that they describe, and develop a Markov chain Monte Carlo scheme for learning annotated hierarchies. We show that our model discov- ers interpretable structure in several real-world data sets.

****

Single Channel Speech Separation Using Factorial Dynamics

John Hershey, Trausti Kristjansson, Steven Rennie, Peder A. Olsen

Human listeners have the extraordinary ability to hear and recognize speech even when more than one person is talking. Their machine counterparts have historically been unable to compete with this ability, until now. We present a modelbased system that performs on par with humans in the task of separating speech of two talkers from a single-channel recording. Remarkably, the system surpasses human recognition performance in many conditions. The models of speech use temporal dynamics to help infer the source speech signals, given mixed speech signals. The estimated source signals are then recognized using a conventional speech recognition system. We demonstrate that the system achieves its best performance when the model of temporal dynamics closely captures the grammatical constraints of the task. One of the hallmarks of human perception is our ability to solve the auditory cocktail party problem: we can direct our attention to a given speaker in the presence of interfering speech, and understand what was said remarkably well. Until now the same could not be said for automatic speech recognition systems. However, we have recently introduced a system which in many conditions performs this task better than humans [1][2]. The model addresses the Pascal Speech Separation Challenge task [3], and outperforms all other published results by more than 10% word error rate (WER). In this model, dynamics are modeled using a layered combination of one or two Markov chains: one for long-term dependencies and another for short-term dependencies. The combination of the two speakers was handled via an iterative Laplace approximation method known as Algonquin [4]. Here we describe experiments that show better performance on the same task with a simpler version of the model. The task we address is provided by the PASCAL Speech Separation Challenge [3], which provides standard training, development, and test data sets of single-channel speech mixtures following an arbitrary but simple grammar. In addition, the challenge organizers have conducted human-listening experiments to provide an interesting baseline for comparison of computational tec

hniques. The overall system we developed is composed of the three components: a speaker identification and gain estimation component, a signal separation component, and a speech recognition system. In this paper we focus on the signal separation component, which is composed of the acoustic and grammatical models. The details of the other components are discussed in [2]. Single-channel speech separation has previously been attempted using Gaussian mixture models (GMMs) on individual frames of acoustic features. However such models tend to perform well only when speakers are of different gender or have rather different voices [4]. When speakers have similar voices, speaker-dependent mixture models cannot unambiguously identify the component speakers. In such cases it is helpful to model the temporal dynamics of the speech. Several models in the literature have attempted to do so either for recognition [5, 6] or enhancement [7, 8] of speech. Such
****

A Small World Threshold for Economic Network Formation

Eyal Even-dar, Michael Kearns

We introduce a game-theoretic model for network formation inspired by earlier stochastic models that mix localized and long-distance connectivity. In this model, players may purchase edges at distance d at a cost of d , and wish to minimize the sum of their edge purchases and their average distance to other players. In this model, we show there is a striking "small world" threshold phenomenon: in two dimensions, if < 2 then every Nash equilibrium results in a network of constant diameter (independent of network size), and if > 2 then every Nash equilibrium results in a network whose diameter grows as a root of the network size, and thus is unbounded. We contrast our results with those of Kleinberg [8] in a stochastic model, and empirically investigate the "navigability" of equilibrium networks. Our theoretical results all generalize to higher dimensions.
****

iLSTD: Eligibility Traces and Convergence Analysis

Alborz Geramifard, Michael Bowling, Martin Zinkevich, Richard S. Sutton

We present new theoretical and empirical results with the iLSTD algorithm for policy evaluation in reinforcement learning with linear function approximation. iLSTD is an incremental method for achieving results similar to LSTD, the dataefficient, least-squares version of temporal difference learning, without incurring the full cost of the LSTD computation. LSTD is O(n2), where n is the number of parameters in the linear function approximator, while iLSTD is O(n). In this paper, we generalize the previous iLSTD algorithm and present three new results: (1) the first convergence proof for an iLSTD algorithm; (2) an extension to incorporate eligibility traces without changing the asymptotic computational complexity; and (3) the first empirical results with an iLSTD algorithm for a problem (mountain car) with feature vectors large enough (n = 10, 000) to show substantial computational advantages over LSTD.
****

Sparse Kernel Orthonormalized PLS for feature extraction in large data sets

Jerónimo Arenas-garcía, Kaare Petersen, Lars K. Hansen

In this paper we are presenting a novel multivariate analysis method. Our scheme is based on a novel kernel orthonormalized partial least squares (PLS) variant for feature extraction, imposing sparsity constrains in the solution to improve scalability. The algorithm is tested on a benchmark of UCI data sets, and on the analysis of integrated short-time music features for genre prediction. The upshot is that the method has strong expressive power even with rather few features, is clearly outperforming the ordinary kernel PLS, and therefore is an appealing method for feature extraction of labelled data.
****

Data Integration for Classification Problems Employing Gaussian Process Priors

Mark Girolami, Mingjun Zhong

By adopting Gaussian process priors a fully Bayesian solution to the problem of integrating possibly heterogeneous data sets within a classification setting is presented. Approximate inference schemes employing Variational & Expectation Propagation based methods are developed and rigorously assessed. We demonstrate our approach to integrating multiple data sets on a large scale protein fold predic tion problem where we infer the optimal combinations of covariance functions and achieve state-of-the-art performance without resorting to any ad hoc parameter tuning and classifier combination.

****

Stochastic Relational Models for Discriminative Link Prediction

Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, Zhao Xu

We introduce a Gaussian process (GP) framework, stochastic relational models (SRM), for learning social, physical, and other relational phenomena where interactions between entities are observed. The key idea is to model the stochastic structure of entity relationships (i.e., links) via a tensor interaction of multiple GPs, each defined on one type of entities. These models in fact define a set of nonparametric priors on infinite dimensional tensor matrices, where each element represents a relationship between a tuple of entities. By maximizing the marginalized likelihood, information is exchanged between the participating GPs through the entire relational network, so that the dependency structure of links is messaged to the dependency of entities, reflected by the adapted GP kernels. The framework offers a discriminative approach to link prediction, namely, predicting the existences, strengths, or types of relationships based on the partially observed linkage network as well as the attributes of entities (if given). We discuss properties and variants of SRM and derive an efficient learning algorithm. Very encouraging experimental results are achieved on a toy problem and a user-movie preference link prediction task. In the end we discuss extensions of SRM to general relational learning tasks.

****

Adaptive Spatial Filters with predefined Region of Interest for EEG based Brain-Computer-Interfaces

Moritz Grosse-wentrup, Klaus Gramann, Martin Buss

The performance of EEG-based Brain-Computer-Interfaces (BCIs) critically depends on the extraction of features from the EEG carrying information relevant for the classification of different mental states. For BCIs employing imaginary movements of different limbs, the method of Common Spatial Patterns (CSP) has been shown to achieve excellent classification results. The CSP-algorithm however suffers from a lack of robustness, requiring training data without artifacts for good performance. To overcome this lack of robustness, we propose an adaptive spatial filter that replaces the training data in the CSP approach by a-priori information. More specifically, we design an adaptive spatial filter that maximizes the ratio of the variance of the electric field originating in a predefined region of interest (ROI) and the overall variance of the measured EEG. Since it is known that the component of the EEG used for discriminating imaginary movements originates in the motor cortex, we design two adaptive spatial filters with the ROIs centered in the hand areas of the left and right motor cortex. We then use these to classify EEG data recorded during imaginary movements of the right and left hand of three subjects, and show that the adaptive spatial filters outperform the CSP-algorithm, enabling classification rates of up to 94.7 % without artifact rejection.

****

A recipe for optimizing a time-histogram

Hideaki Shimazaki, Shigeru Shinomoto

The time-histogram method is a handy tool for capturing the instantaneous rate of spike occurrence. In most of the neurophysiological literature, the bin size that critically determines the goodness of the fit of the time-histogram to the underlying rate has been selected by individual researchers in an unsystematic manner. We propose an objective method for selecting the bin size of a time-histogram from the spike data, so that the time-histogram best approximates the unknown underlying rate. The resolution of the histogram increases, or the optimal bin size decreases, with the number of spike sequences sampled. It is notable that the optimal bin size diverges if only a small number of experimental trials are available from a moderately fluctuating rate process. In this case, any attempt to characterize the underlying spike rate will lead to spurious results. Given a paucity of data, our method can also suggest how many more trials are needed un til the set of data can be analyzed with the required resolution.

****

Inducing Metric Violations in Human Similarity Judgements

Julian Laub, Klaus-Robert Müller, Felix A. Wichmann, Jakob H. Macke

Attempting to model human categorization and similarity judgements is both a very interesting but also an exceedingly difficult challenge. Some of the difficulty arises because of conflicting evidence whether human categorization and similarity judgements should or should not be modelled as to operate on a mental representation that is essentially metric. Intuitively, this has a strong appeal as it would allow (dis)similarity to be represented geometrically as distance in some internal space. Here we show how a single stimulus, carefully constructed in a psychophysical experiment, introduces l2 violations in what used to be an internal similarity space that could be adequately modelled as Euclidean. We term this one influential data point a conflictual judgement. We present an algorithm of how to analyse such data and how to identify the crucial point. Thus there may not be a strict dichotomy between either a metric or a non-metric internal space but rather degrees to which potentially large subsets of stimuli are represented metrically with a small subset causing a global violation of metricity.

****

Nonnegative Sparse PCA

Ron Zass, Amnon Shashua

We describe a nonnegative variant of the "Sparse PCA" problem. The goal is to create a low dimensional representation from a collection of points which on the one hand maximizes the variance of the projected points and on the other uses only parts of the original coordinates, and thereby creating a sparse representation. What distinguishes our problem from other Sparse PCA formulations is that the projection involves only nonnegative weights of the original coordinates -- a desired quality in various fields, including economics, bioinformatics and computer vision. Adding nonnegativity contributes to sparseness, where it enforces a partitioning of the original coordinates among the new axes. We describe a simple yet efficient iterative coordinate-descent type of scheme which converges to a local optimum of our optimization criteria, giving good results on large real world datasets.

****

Context dependent amplification of both rate and event-correlation in a VLSI network of spiking neurons

Elisabetta Chicca, Giacomo Indiveri, Rodney Douglas

Cooperative competitive networks are believed to play a central role in cortical processing and have been shown to exhibit a wide set of useful computational properties. We propose a VLSI implementation of a spiking cooperative competitive network and show how it can perform context dependent computation both in the mean firing rate domain and in spike timing correlation space. In the mean rate case the network amplifies the activity of neurons belonging to the selected stimulus and suppresses the activity of neurons receiving weaker stimuli. In the event correlation case, the recurrent network amplifies with a higher gain the correlation between neurons which receive highly correlated inputs while leaving the mean firing rate unaltered. We describe the network architecture and present experimental data demonstrating its context dependent computation capabilities.

****

PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, Nicolas Usunier

We propose new PAC-Bayes bounds for the risk of the weighted majority vote that depend on the mean and variance of the error of its associated Gibbs classifier. We show that these bounds can be smaller than the risk of the Gibbs classifier and can be arbitrarily close to zero even if the risk of the Gibbs classifier is close to 1/2. Moreover, we show that these bounds can be uniformly estimated on the training data for all possible posteriors Q. Moreover, they can be improved by using a large sample of unlabelled data.

****

Generalized Regularized Least-Squares Learning with Predefined Features in a Hilbert Space

Wenye Li, Kin-hong Lee, Kwong-sak Leung

Kernel-based regularized learning seeks a model in a hypothesis space by minimizing the empirical error and the model's complexity. Based on the representer theorem, the solution consists of a linear combination of translates of a kernel. This paper investigates a generalized form of representer theorem for kernel-based learning. After mapping predefined features and translates of a kernel simultaneously onto a hypothesis space by a specific way of constructing kernels, we proposed a new algorithm by utilizing a generalized regularizer which leaves part of the space unregularized. Using a squared-loss function in calculating the empirical error, a simple convex solution is obtained which combines predefined features with translates of the kernel. Empirical evaluations have confirmed the effectiveness of the algorithm for supervised learning tasks.

****

Map-Reduce for Machine Learning on Multicore

Cheng-tao Chu, Sang Kim, Yi-an Lin, Yuanyuan Yu, Gary Bradski, Kunle Olukotun, Andrew Ng

We are at the beginning of the multicore era. Computers will have increasingly many cores (processors), but there is still no good programming framework for these architectures, and thus no simple and unified way for machine learning to take advantage of the potential speed up. In this paper, we develop a broadly applicable parallel programming method, one that is easily applied to many different learning algorithms. Our work is in distinct contrast to the tradition in machine learning of designing (often ingenious) ways to speed up a single algorithm at a time. Specifically, we show that algorithms that fit the Statistical Query model [15] can be written in a certain "summation form," which allows them to be easily parallelized on multicore computers. We adapt Google's map-reduce [7] paradigm to demonstrate this parallel speed up technique on a variety of learning algorithms including locally weighted linear regression (LWLR), k-means, logistic regression (LR), naive Bayes (NB), SVM, ICA, PCA, gaussian discriminant analysis (GDA), EM, and backpropagation (NN). Our experimental results show basically linear speedup with an increasing number of processors.

****

Natural Actor-Critic for Road Traffic Optimisation

Silvia Richter, Douglas Aberdeen, Jin Yu

Current road-traffic optimisation practice around the world is a combination of hand tuned policies with a small degree of automatic adaption. Even state-ofthe-art research controllers need good models of the road traffic, which cannot be obtained directly from existing sensors. We use a policy-gradient reinforcement learning approach to directly optimise the traffic signals, mapping currently deployed sensor observations to control signals. Our trained controllers are (theoretically) compatible with the traffic system used in Sydney and many other cities around the world. We apply two policy-gradient methods: (1) the recent natural actor-critic algorithm, and (2) a vanilla policy-gradient algorithm for comparison. Along the way we extend natural-actor critic approaches to work for distributed and online infinite-horizon problems.

****

Implicit Surfaces with Globally Regularised and Compactly Supported Basis Functions

Christian Walder, Olivier Chapelle, Bernhard Schölkopf

We consider the problem of constructing a function whose zero set is to represent a surface, given sample points with surface normal vectors. The contributions include a novel means of regularising multi-scale compactly supported basis functions that leads to the desirable properties previously only associated with fully supported bases, and show equivalence to a Gaussian process with modified covariance function. We also provide a regularisation framework for simpler and more direct treatment of surface normals, along with a corresponding generalisation of the representer theorem. We demonstrate the techniques on 3D problems of up to 14 million data points, as well as 4D time series data.

****

Multiple Instance Learning for Computer Aided Diagnosis
Murat Dundar, Balaji Krishnapuram, R. Rao, Glenn Fung

Many computer aided diagnosis (CAD) problems can be best modelled as a multiple-instance learning (MIL) problem with unbalanced data: i.e. , the training data typically consists of a few positive bags, and a very large number of negative instances. Existing MIL algorithms are much too computationally expensive for these datasets. We describe CH, a framework for learning a Convex Hull representation of multiple instances that is significantly faster than existing MIL algorithms. Our CH framework applies to any standard hyperplane-based learning algorithm, and for some algorithms, is guaranteed to find the global optimal solution. Experimental studies on two different CAD applications further demonstrate that the proposed algorithm significantly improves diagnostic accuracy when compared to both MIL and traditional classifiers. Although not designed for standard MIL problems (which have both positive and negative bags and relatively balanced datasets), comparisons against other MIL methods on benchmark problems also indicate that the proposed method is competitive with the state-of-the-art.

****

A selective attention multi--chip system with dynamic synapses and spiking neurons
Chiara Bartolozzi, Giacomo Indiveri

Selective attention is the strategy used by biological sensory systems to solve the problem of limited parallel processing capacity: salient subregions of the input stimuli are serially processed, while nonsalient regions are suppressed. We present an mixed mode analog/digital Very Large Scale Integration implementation of a building block for a multichip neuromorphic hardware model of selective attention. We describe the chip's architecture and its behavior, when its is part of a multichip system with a spiking retina as input, and show how it can be used to implement in real-time flexible models of bottom-up attention.

****

Real-time adaptive information-theoretic optimization of neurophysiology experiments
Jeremy Lewi, Robert Butera, Liam Paninski

Adaptively optimizing experiments can significantly reduce the number of trials needed to characterize neural responses using parametric statistical models. However, the potential for these methods has been limited to date by severe computational challenges: choosing the stimulus which will provide the most information about the (typically high-dimensional) model parameters requires evaluating a high-dimensional integration and optimization in near-real time. Here we present a fast algorithm for choosing the optimal (most informative) stimulus based on a Fisher approximation of the Shannon information and specialized numerical linear algebra techniques. This algorithm requires only low-rank matrix manipulations and a one-dimensional linesearch to choose the stimulus and is therefore efficient even for high-dimensional stimulus and parameter spaces; for example, we require just 15 milliseconds on a desktop computer to optimize a 100-dimensional stimulus. Our algorithm therefore makes real-time adaptive experimental design feasible. Simulation results show that model parameters can be estimated much more efficiently using these adaptive techniques than by using random (nonadaptive) stimuli. Finally, we generalize the algorithm to efficiently handle both fast adaptation due to spike-history effects and slow, non-systematic drifts in the model parameters. Maximizing the efficiency of data collection is important in any experimental setting. In neurophysiology experiments, minimizing the number of trials needed to characterize a neural system is essential for maintaining the viability of a preparation and ensuring robust results. As a result, various approaches have been developed to optimize neurophysiology experiments online in order to choose the "best" stimuli given prior knowledge of the system and the observed history of the cell's responses. The "best" stimulus can be defined a number of different ways depending on the experimental objectives. One reasonable choice, if we are interested in finding a neuron's "preferred stimulus," is the stimu lus which maximizes the firing rate of the neuron [1, 2, 3, 4]. Alternatively, when investigating the coding properties of sensory cells it makes sense to define the optimal stimulus in terms of the mutual information between the stimulus and response [5]. Here we take a system identification approach: we define the optimal stimulus as the one which tells us the most about how a neural system responds to its inputs [6, 7]. We consider neural systems in

****

High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression

Martin J. Wainwright, John Lafferty, Pradeep Ravikumar

We focus on the problem of estimating the graph structure associated with a discrete Markov random field. We describe a method based on 1- regularized logistic regression, in which the neighborhood of any given node is estimated by performing logistic regression subject to an1-constraint. Our framework applies to the high-dimensional setting, in which both the number of nodes p and maximum neighborhood sizes d are allowed to grow as a function of the number of observations n. Our main result is to estab- lish sufficient conditions on the triple (n, p, d) for the method to succeed in consistently estimating the neighborhood of every node in the graph simul- taneously. Under certain mutual incoherence conditions analogous to those imposed in previous work on linear regression, we prove that consistent neighborhood selection can be obtained as long as the number of observa- tions n grows more quickly than 6d6 log d + 2d5 log p, thereby establishing that logarithmic growth in the number of samples n relative to graph size p is sufficient to achieve neighborhood consistency.

****

Efficient Learning of Sparse Representations with an Energy-Based Model

Marc'aurelio Ranzato, Christopher Poultney, Sumit Chopra, Yann Cun

We describe a novel unsupervised method for learning sparse, overcomplete features. The model uses a linear encoder, and a linear decoder preceded by a sparsifying non-linearity that turns a code vector into a quasi-binary sparse code vector. Given an input, the optimal code minimizes the distance between the output of the decoder and the input patch while being as similar as possible to the encoder output. Learning proceeds in a two-phase EM-like fashion: (1) compute the minimum-energy code vector, (2) adjust the parameters of the encoder and decoder so as to decrease the energy. The model produces "stroke detectors" when trained on handwritten numerals, and Gabor-like filters when trained on natural image patches. Inference and learning are very fast, requiring no preprocessing, and no expensive sampling. Using the proposed unsupervised method to initialize the first layer of a convolutional network, we achieved an error rate slightly lower than the best reported result on the MNIST dataset. Finally, an extension of the method is described to learn topographical filter maps.

****

Learning Time-Intensity Profiles of Human Activity using Non-Parametric Bayesian Models

Alexander Ihler, Padhraic Smyth

Data sets that characterize human activity over time through collections of timestamped events or counts are of increasing interest in application areas as humancomputer interaction, video surveillance, and Web data analysis. We propose a non-parametric Bayesian framework for modeling collections of such data. In particular, we use a Dirichlet process framework for learning a set of intensity functions corresponding to different categories, which form a basis set for representing individual time-periods (e.g., several days) depending on which categories the time-periods are assigned to. This allows the model to learn in a data-driven fashion what "factors" are generating the observations on a particular day, including (for example) weekday versus weekend effects or day-specific effects corresponding to unique (single-day) occurrences of unusual behavior, sharing information where appropriate to obtain improved estimates of the behavior associated with each category. Applications to realworld data sets of count data involving both vehicles and people are used to illustrate the technique.

****

Using Combinatorial Optimization within Max-Product Belief Propagation

Daniel Tarlow, Gal Elidan, Daphne Koller, John C. Duchi

In general, the problem of computing a maximum a posteriori (MAP) assignment in a Markov random field (MRF) is computationally intractable. However, in certain subclasses of MRF, an optimal or close-to-optimal assignment can be found very efficiently using combinatorial optimization algorithms: certain MRFs with mutual exclusion constraints can be solved using bipartite matching, and MRFs with regular potentials can be solved using minimum cut methods. However, these solutions do not apply to the many MRFs that contain such tractable components as sub-networks, but also other non-complying potentials. In this paper, we present a new method, called COMPOSE, for exploiting combinatorial optimization for sub-networks within the context of a max-product belief propagation algorithm. COMPOSE uses combinatorial optimization for computing exact maxmarginals for an entire sub-network; these can then be used for inference in the context of the network as a whole. We describe highly efficient methods for computing max-marginals for subnetworks corresponding both to bipartite matchings and to regular networks. We present results on both synthetic and real networks encoding correspondence problems between images, which involve both matching constraints and pairwise geometric constraints. We compare to a range of current methods, showing that the ability of COMPOSE to transmit information globally across the network leads to improved convergence, decreased running time, and higher-scoring assignments.

****

Branch and Bound for Semi-Supervised Support Vector Machines

Olivier Chapelle, Vikas Sindhwani, S. Keerthi

Semi-supervised SVMs (S3 VM) attempt to learn low-density separators by maximizing the margin over labeled and unlabeled examples. The associated optimization problem is non-convex. To examine the full potential of S3 VMs modulo local minima problems in current implementations, we apply branch and bound techniques for obtaining exact, global ly optimal solutions. Empirical evidence suggests that the globally optimal solution can return excellent generalization performance in situations where other implementations fail completely. While our current implementation is only applicable to small datasets, we discuss variants that can potentially lead to practically useful algorithms.

****

Differential Entropic Clustering of Multivariate Gaussians

Jason Davis, Inderjit Dhillon

Gaussian data is pervasive and many learning algorithms (e.g., k -means) model their inputs as a single sample drawn from a multivariate Gaussian. However, in many real-life settings, each input object is best described by multiple samples drawn from a multivariate Gaussian. Such data can arise, for example, in a movie review database where each movie is rated by several users, or in time-series domains such as sensor networks. Here, each input can be naturally described by both a mean vector and covariance matrix which parameterize the Gaussian distribution. In this paper, we consider the problem of clustering such input objects, each represented as a multivariate Gaussian. We formulate the problem using an information theoretic approach and draw several interesting theoretical connections to Bregman divergences and also Bregman matrix divergences. We evaluate our method across several domains, including synthetic data, sensor network data, and a statistical debugging application.

****

Multi-Instance Multi-Label Learning with Application to Scene Classification

Zhi-Li Zhang, Min-ling Zhang

In this paper, we formalize multi-instance multi-label learning, where each train- ing example is associated with not only multiple instances but also multiple class labels. Such a problem can occur in many real-world tasks, e.g. an image usually contains multiple patches each of which can be described by a feature vector, and the image can belong to multiple categories since its semantics can be recognized in different ways. We analyze the relationship between multi-instance multi-label learning and the learning frameworks of traditional supervised lear ning, multi- instance learning and multi-label learning. Then, we propose the MIMLBOOST and MIMLSVM algorithms which achieve good performance in an application to scene classification.

****

Analysis of Contour Motions

Ce Liu, William Freeman, Edward Adelson

A reliable motion estimation algorithm must function under a wide range of conditions. One regime, which we consider here, is the case of moving objects with contours but no visible texture. Tracking distinctive features such as corners can disambiguate the motion of contours, but spurious features such as T-junctions can be badly misleading. It is difficult to determine the reliability of motion from local measurements, since a full rank covariance matrix can result from both real and spurious features. We propose a novel approach that avoids these points al- together, and derives global motion estimates by utilizing information from three levels of contour analysis: edgelets, boundary fragments and contours. Boundary fragment are chains of orientated edgelets, for which we derive motion estimates from local evidence. The uncertainties of the local estimates are disambiguated after the boundary fragments are properly grouped into contours. The grouping is done by constructing a graphical model and marginalizing it using importance sampling. We propose two equivalent representations in this graphical model, re- versible switch variables attached to the ends of fragments and fragment chains, to capture both local and global statistics of boundaries. Our system is success- fully applied to both synthetic and real video sequences containing high-contrast boundaries and textureless regions. The system produces good motion estimates along with properly grouped and completed contours.

****

Learning to be Bayesian without Supervision

Martin Raphan, Eero Simoncelli

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

****

Causal inference in sensorimotor integration

Konrad Körding, Joshua Tenenbaum

Many recent studies analyze how data from different modalities can be combined. Often this is modeled as a system that optimally combines several sources of information about the same variable. However, it has long been realized that this information combining depends on the interpretation of the data. Two cues that are perceived by different modalities can have different causal relationships: (1) They can both have the same cause, in this case we should fully integrate both cues into a joint estimate. (2) They can have distinct causes, in which case information should be processed independently. In many cases we will not know if there is one joint cause or two independent causes that are responsible for the cues. Here we model this situation as a Bayesian estimation problem. We are thus able to explain some experiments on visual auditory cue combination as well as some experiments on visual proprioceptive cue integration. Our analysis shows that the problem solved by people when they combine cues to produce a movement is much more complicated than is usually assumed, because they need to infer the causal structure that is underlying their sensory experience.

****

A Kernel Subspace Method by Stochastic Realization for Learning Nonlinear Dynamical Systems

Yoshinobu Kawahara, Takehisa Yairi, Kazuo Machida

In this paper, we present a subspace method for learning nonlinear dynamical systems based on stochastic realization, in which state vectors are chosen using kernel canonical correlation analysis, and then state-space systems are identified through regression with the state vectors. We construct the theoretical underpinning and derive a concrete algorithm for nonlinear identification. The obtained algorithm needs no iterative optimization procedure and can be implemented on t he basis of fast and reliable numerical schemes. The simulation result shows that our algorithm can express dynamics with a high degree of accuracy.

****

Approximate inference using planar graph decomposition

Amir Globerson, Tommi Jaakkola

A number of exact and approximate methods are available for inference calculations in graphical models. Many recent approximate methods for graphs with cycles are based on tractable algorithms for tree structured graphs. Here we base the approximation on a different tractable model, planar graphs with binary variables and pure interaction potentials (no external field). The partition function for such models can be calculated exactly using an algorithm introduced by Fisher and Kasteleyn in the 1960s. We show how such tractable planar models can be used in a decomposition to derive upper bounds on the partition function of non-planar models. The resulting algorithm also allows for the estimation of marginals. We compare our planar decomposition to the tree decomposition method of Wainwright et. al., showing that it results in a much tighter bound on the partition function, improved pairwise marginals, and comparable singleton marginals. Graphical models are a powerful tool for modeling multivariate distributions, and have been successfully applied in various fields such as coding theory and image processing. Applications of graphical models typically involve calculating two types of quantities, namely marginal distributions, and MAP assignments. The evaluation of the model partition function is closely related to calculating marginals [12]. These three problems can rarely be solved exactly in polynomial time, and are provably computationally hard in the general case [1]. When the model conforms to a tree structure, however, all these problems can be solved in polynomial time. This has prompted extensive research into tree based methods. For example, the junction tree method [6] converts a graphical model into a tree by clustering nodes into cliques, such that the graph over cliques is a tree. The resulting maximal clique size (cf. tree width) may nevertheless be prohibitively large. Wainwright et. al. [9, 11] proposed an approximate method based on trees known as tree reweighting (TRW). The TRW approach decomposes the potential vector of a graphical model into a mixture over spanning trees of the model, and then uses convexity arguments to bound various quantities, such as the partition function. One key advantage of this approach is that it provides bounds on partition function value, a property which is not shared by approximations based on Bethe free energies [13]. In this paper we focus on a different class of tractable models: planar graphs. A graph is called planar if it can be drawn in the plane without crossing edges. Works in the 1960s by physicists Fisher [5] and Kasteleyn [7], among others, have shown that the partition function for planar graphs may be calculated in polynomial time. This, however, is true under two key restrictions. One is that the variables xi are binary. The other is that the interaction potential depends only on xi xj (where xi {1}), and not on their individual values (i.e., the zero external field case). Here we show how the above method can be used to obtain upper bounds on the partition function for non-planar graphs. As in TRW, we decompose the potential of a non-planar graph into a sum

****

Context Effects in Category Learning: An Investigation of Four Probabilistic Models

Michael C. Mozer, Michael Shettel, Michael Holmes

Categorization is a central activity of human cognition. When an individual is asked to categorize a sequence of items, context effects arise: categorization of one item influences category decisions for subsequent items. Specifically, when experimental subjects are shown an exemplar of some target category, the category prototype appears to be pulled toward the exemplar, and the prototypes of all nontarget categories appear to be pushed away. These push and pull effects diminish with experience, and likely reflect long-term learning of category boundaries. We propose and evaluate four principled probabilistic (Bayesian) accounts of context effects in categorization. In all four accounts, the probability of an exemplar given a category is encoded as a Gaussian density in feature space, and categorization involves computing category posteriors given an exemplar. The mo dels differ in how the uncertainty distribution of category prototypes is represented (localist or distributed), and how it is updated following each experience (using a maximum likelihood gradient ascent, or a Kalman filter update). We find that the distributed maximum-likelihood model can explain the key experimental phenomena. Further, the model predicts other phenomena that were confirmed via reanalysis of the experimental data.

****

An Application of Reinforcement Learning to Aerobatic Helicopter Flight

Pieter Abbeel, Adam Coates, Morgan Quigley, Andrew Ng

Autonomous helicopter flight is widely regarded to be a highly challenging control problem. This paper presents the first successful autonomous completion on a real RC helicopter of the following four aerobatic maneuvers: forward flip and sideways roll at low speed, tail-in funnel, and nose-in funnel. Our experimental results significantly extend the state of the art in autonomous helicopter flight. We used the following approach: First we had a pilot fly the helicopter to help us find a helicopter dynamics model and a reward (cost) function. Then we used a reinforcement learning (optimal control) algorithm to find a controller that is optimized for the resulting model and reward function. More specifically, we used differential dynamic programming (DDP), an extension of the linear quadratic regulator (LQR).

****

Parameter Expanded Variational Bayesian Methods

Tommi Jaakkola, Yuan Qi

Bayesian inference has become increasingly important in statistical machine learning. Exact Bayesian calculations are often not feasible in practice, however. A number of approximate Bayesian methods have been proposed to make such calculations practical, among them the variational Bayesian (VB) approach. The VB approach, while useful, can nevertheless suffer from slow convergence to the approximate solution. To address this problem, we propose Parameter-eXpanded Variational Bayesian (PX-VB) methods to speed up VB. The new algorithm is inspired by parameter-expanded expectation maximization (PX-EM) and parameterexpanded data augmentation (PX-DA). Similar to PX-EM and -DA, PX-VB expands a model with auxiliary variables to reduce the coupling between variables in the original model. We analyze the convergence rates of VB and PX-VB and demonstrate the superior convergence rates of PX-VB in variational probit regression and automatic relevance determination.

****

An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments

Michael Mandel, Daniel Ellis, Tony Jebara

We present a method for localizing and separating sound sources in stereo recordings that is robust to reverberation and does not make any assumptions about the source statistics. The method consists of a probabilistic model of binaural multisource recordings and an expectation maximization algorithm for finding the maximum likelihood parameters of that model. These parameters include distributions over delays and assignments of time-frequency regions to sources. We evaluate this method against two comparable algorithms on simulations of simultaneous speech from two or three sources. Our method outperforms the others in anechoic conditions and performs as well as the better of the two in the presence of reverberation.

****

Recursive ICA

Honghao Shan, Lingyun Zhang, Garrison Cottrell

Independent Component Analysis (ICA) is a popular method for extracting independent features from visual data. However, as a fundamentally linear technique, there is always nonlinear residual redundancy that is not captured by ICA. Hence there have been many attempts to try to create a hierarchical version of ICA, but so far none of the approaches have a natural way to apply them more than once. Here we show that there is a relatively simple technique that transforms the absolute values of the outputs of a previous application of ICA into a normal distri bution, to which ICA maybe applied again. This results in a recursive ICA algorithm that may be applied any number of times in order to extract higher order structure from previous layers.

****

The Neurodynamics of Belief Propagation on Binary Markov Random Fields

Thomas Ott, Ruedi Stoop

We rigorously establish a close relationship between message passing algorithms and models of neurodynamics by showing that the equations of a continuous Hop- (cid:2)eld network can be derived from the equations of belief propagation on a binary Markov random (cid:2)eld. As Hop(cid:2)eld networks are equipped with a Lyapunov func- tion, convergence is guaranteed. As a consequence, in the limit of many weak con- nections per neuron, Hop(cid:2)eld networks exactly implement a continuous-time vari- ant of belief propagation starting from message initialisations that prevent from running into convergence problems. Our results lead to a better understanding of the role of message passing algorithms in real biological neural networks.

****

Unsupervised Learning of a Probabilistic Grammar for Object Detection and Parsing

Yuanhao Chen, Long Zhu, Alan L. Yuille

We describe an unsupervised method for learning a probabilistic grammar of an object from a set of training examples. Our approach is invariant to the scale and rotation of the objects. We illustrate our approach using thirteen objects from the Caltech 101 database. In addition, we learn the model of a hybrid object class where we do not know the specific object or its position, scale or pose. This is illustrated by learning a hybrid class consisting of faces, motorbikes, and airplanes. The individual objects can be recovered as different aspects of the grammar for the object class. In all cases, we validate our results by learning the probability grammars from training datasets and evaluating them on the test datasets. We compare our method to alternative approaches. The advantages of our approach is the speed of inference (under one second), the parsing of the object, and increased accuracy of performance. Moreover, our approach is very general and can be applied to a large range of objects and structures.

****

A Humanlike Predictor of Facial Attractiveness

Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-or, Eytan Ruppin

This work presents a method for estimating human facial attractiveness, based on supervised learning techniques. Numerous facial features that describe facial geometry, color and texture, combined with an average human attractiveness score for each facial image, are used to train various predictors. Facial attractiveness ratings produced by the final predictor are found to be highly correlated with human ratings, markedly improving previous machine learning achievements. Simulated psychophysical experiments with virtually manipulated images reveal preferences in the machine's judgments which are remarkably similar to those of humans. These experiments shed new light on existing theories of facial attractiveness such as the averageness, smoothness and symmetry hypotheses. It is intriguing to find that a machine trained explicitly to capture an operational performance criteria such as attractiveness rating, implicitly captures basic human psychophysical biases characterizing the perception of facial attractiveness in general.

****

Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype

Tobias Sing, Niko Beerenwinkel

Starting with the work of Jaakkola and Haussler, a variety of approaches have been proposed for coupling domain-specific generative models with statistical learning methods. The link is established by a kernel function which provides a similarity measure based inherently on the underlying model. In computational biology, the full promise of this framework has rarely ever been exploited, as most kernels are derived from very generic models, such as sequence profiles or hidden Markov models. Here, we introduce the MTreeMix kernel, which is based on a gener ative model tailored to the underlying biological mechanism. Specifically, the kernel quantifies the similarity of evolutionary escape from antiviral drug pressure between two viral sequence samples. We compare this novel kernel to a standard, evolution-agnostic amino acid encoding in the prediction of HIV drug resistance from genotype, using support vector regression. The results show significant improvements in predictive performance across 17 anti-HIV drugs. Thus, in our study, the generative-discriminative paradigm is key to bridging the gap between population genetic modeling and clinical decision making.

****

Image Retrieval and Classification Using Local Distance Functions

Andrea Frome, Yoram Singer, Jitendra Malik

In this paper we introduce and experiment with a framework for learning local perceptual distance functions for visual recognition. We learn a distance function for each training image as a combination of elementary distances between patch-based visual features. We apply these combined local distance functions to the tasks of image retrieval and classification of novel images. On the Caltech 101 object recognition benchmark, we achieve 60.3% mean recognition across classes using 15 training images per class, which is better than the best published performance by Zhang, et al.

****

Chained Boosting

Christian Shelton, Wesley Huie, Kin Kan

We describe a method to learn to make sequential stopping decisions, such as those made along a processing pipeline. We envision a scenario in which a series of decisions must be made as to whether to continue to process. Further processing costs time and resources, but may add value. Our goal is to create, based on his- toric data, a series of decision rules (one at each stage in the pipeline) that decide, based on information gathered up to that point, whether to continue processing the part. We demonstrate how our framework encompasses problems from manu- facturing to vision processing. We derive a quadratic (in the number of decisions) bound on testing performance and provide empirical results on object detection.

****

Support Vector Machines on a Budget

Ofer Dekel, Yoram Singer

The standard Support Vector Machine formulation does not provide its user with the ability to explicitly control the number of support vectors used to define the generated classifier. We present a modified version of SVM that allows the user to set a budget parameter B and focuses on minimizing the loss attained by the B worst-classified examples while ignoring the remaining examples. This idea can be used to derive sparse versions of both L1-SVM and L2-SVM. Technically, we obtain these new SVM variants by replacing the 1-norm in the standard SVM for- mulation with various interpolation-norms. We also adapt the SMO optimization algorithm to our setting and report on some preliminary experimental results.

****

Manifold Denoising

Matthias Hein, Markus Maier

We consider the problem of denoising a noisily sampled submanifold M in Rd, where the submanifold M is a priori unknown and we are only given a noisy point sample. The presented denoising algorithm is based on a graph-based diffusion process of the point sample. We analyze this diffusion process using recent re- sults about the convergence of graph Laplacians. In the experiments we show that our method is capable of dealing with non-trivial high-dimensional noise. More- over using the denoising algorithm as pre-processing method we can improve the results of a semi-supervised learning algorithm.

****

Learning to Model Spatial Dependency: Semi-Supervised Discriminative Random Fields

Chi-hoon Lee, Shaojun Wang, Feng Jiao, Dale Schuurmans, Russell Greiner

We present a novel, semi-supervised approach to training discriminative random f ields (DRFs) that efficiently exploits labeled and unlabeled training data to achieve improved accuracy in a variety of image processing tasks. We formulate DRF training as a form of MAP estimation that combines conditional loglikelihood on labeled data, given a data-dependent prior, with a conditional entropy regularizer defined on unlabeled data. Although the training objective is no longer concave, we develop an efficient local optimization procedure that produces classifiers that are more accurate than ones based on standard supervised DRF training. We then apply our semi-supervised approach to train DRFs to segment both synthetic and real data sets, and demonstrate significant improvements over supervised DRFs in each case.

****

Shifting, One-Inclusion Mistake Bounds and Tight Multiclass Expected Risk Bounds

Benjamin Rubinstein, Peter Bartlett, J. Rubinstein

Under the prediction model of learning, a prediction strategy is presented with an i.i.d. sample of n - 1 points in X and corresponding labels from a concept f F , and aims to minimize the worst-case probability of erring on an nth point. By exploiting the structure of F , Haussler et al. achieved a VC(F)/n bound for the natural one-inclusion prediction strategy, improving on bounds implied by PAC-type results by a O(log n) factor. The key data structure in their result is the natural subgraph of the hypercube--the one-inclusion graph; the key step is a d = VC(F) bound on one-inclunion graph density. The first main result of this s /n -1 paper is a density bound of n d-1 (d) < d, which positively resolves a conjecture of Kuzmin & Warmuth relating to their unlabeled Peeling compression scheme and also leads to an improved mistake bound for the randomized (deterministic) one-inclusion strategy for all d (for d (n)). The proof uses a new form of VC-invariant shifting and a group-theoretic symmetrization. Our second main result is a k -class analogue of the d/n mistake bound, replacing the VC-dimension by the Pollard pseudo-dimension and the one-inclusion strategy by its natural hypergraph generalization. This bound on expected risk improves on known PAC-based results by a factor of O(log n) and is shown to be optimal up to a O(log k) factor. The combinatorial technique of shifting takes a central role in understanding the one-inclusion (hyper)graph and is a running theme throughout.

****

Learning to parse images of articulated bodies

Deva Ramanan

We consider the machine vision task of pose estimation from static images, specifically for the case of articulated objects. This problem is hard because of the large number of degrees of freedom to be estimated. Following a established line of research, pose estimation is framed as inference in a probabilistic model. In our experience however, the success of many approaches often lie in the power of the features. Our primary contribution is a novel casting of visual inference as an iterative parsing process, where one sequentially learns better and better features tuned to a particular image. We show quantitative results for human pose estimation on a database of over 300 images that suggest our algorithm is competitive with or surpasses the state-of-the-art. Since our procedure is quite general (it does not rely on face or skin detection), we also use it to estimate the poses of horses in the Weizmann database.

****

Correcting Sample Selection Bias by Unlabeled Data

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, Alex Smola

We consider the scenario where training and test data are drawn from different distributions, commonly referred to as sample selection bias. Most algorithms for this setting try to first recover sampling distributions and then make appro- priate corrections based on the distribution estimate. We present a nonparametric method which directly produces resampling weights without distribution estima- tion. Our method works by matching distributions between training and testing sets in feature space. Experimental results demonstrate that our method works well in practice.

****

A Nonparametric Approach to Bottom-Up Visual Saliency

Wolf Kienzle, Felix A. Wichmann, Matthias Franz, Bernhard Schölkopf

This paper addresses the bottom-up influence of local image information on human eye movements. Most existing computational models use a set of biologically plausible linear filters, e.g., Gabor or Difference-of-Gaussians filters as a front-end, the outputs of which are nonlinearly combined into a real number that indicates visual saliency. Unfortunately, this requires many design parameters such as the number, type, and size of the front-end filters, as well as the choice of nonlinearities, weighting and normalization schemes etc., for which biological plausibility cannot always be justified. As a result, these parameters have to be chosen in a more or less ad hoc way. Here, we propose to learn a visual saliency model directly from human eye movement data. The model is rather simplistic and essentially parameter-free, and therefore contrasts recent developments in the field that usually aim at higher prediction rates at the cost of additional parameters and increasing model complexity. Experimental results show that--despite the lack of any biological prior knowledge--our model performs comparably to existing approaches, and in fact learns image features that resemble findings from several previous studies. In particular, its maximally excitatory stimuli have center-surround structure, similar to receptive fields in the early human visual system.

****

Computation of Similarity Measures for Sequential Data using Generalized Suffix Trees

Konrad Rieck, Pavel Laskov, Sören Sonnenburg

We propose a generic algorithm for computation of similarity measures for se- quential data. The algorithm uses generalized suf■x trees for ef■cient calculation of various kernel, distance and non-metric similarity functions. Its worst-case run-time is linear in the length of sequences and independent of the underlying embedding language, which can cover words, k-grams or all contained subse- quences. Experiments with network intrusion detection, DNA analysis and text processing applications demonstrate the utility of distances and similarity coef■- cients for sequences as alternatives to classical kernel functions.

****

Randomized PCA Algorithms with Regret Bounds that are Logarithmic in the Dimension

Manfred K. K. Warmuth, Dima Kuzmin

We design an on-line algorithm for Principal Component Analysis. In each trial the current instance is projected onto a probabilistically chosen low dimensional subspace. The total expected quadratic approximation error equals the total quadratic approximation error of the best subspace chosen in hindsight plus some additional term that grows linearly in dimension of the subspace but logarithmically in the dimension of the instances.

****

Efficient Structure Learning of Markov Networks using L_1-Regularization

Su-in Lee, Varun Ganapathi, Daphne Koller

Markov networks are commonly used in a wide variety of applications, ranging from computer vision, to natural language, to computational biology. In most current applications, even those that rely heavily on learned models, the structure of the Markov network is constructed by hand, due to the lack of effective algorithms for learning Markov network structure from data. In this paper, we provide a computationally efficient method for learning Markov network structure from data. Our method is based on the use of L1 regularization on the weights of the log-linear model, which has the effect of biasing the model towards solutions where many of the parameters are zero. This formulation converts the Markov network learning problem into a convex optimization problem in a continuous space, which can be solved using efficient gradient methods. A key issue in this setting is the (unavoidable) use of approximate inference, which can lead to errors in the gradient computation when the network structure is dense. Thus, we explore the use of different feature introduction schemes and compare their performance. We provide results for our method on synthetic data, and on two real world data sets: pixel values in the MNIST data, and genetic sequence variations in the human Ha pMap data. We show that our L1 -based method achieves considerably higher generalization performance than the more standard L2 -based method (a Gaussian parameter prior) or pure maximum-likelihood learning. We also show that we can learn MRF network structure at a computational cost that is not much greater than learning parameters alone, demonstrating the existence of a feasible method for this important problem.

****

Attribute-efficient learning of decision lists and linear threshold functions under unconcentrated distributions

Philip Long, Rocco Servedio

We consider the well-studied problem of learning decision lists using few examples when many irrelevant features are present. We show that smooth boosting algorithms such as MadaBoost can efficiently learn decision lists of length k over n boolean variables using poly(k , log n) many examples provided that the marginal distribution over the relevant variables is "not too concentrated" in an L 2 -norm sense. Using a recent result of Hastad, we extend the analysis to obtain a similar (though quantitatively weaker) result for learning arbitrary linear threshold functions with k nonzero coefficients. Experimental results indicate that the use of a smooth boosting algorithm, which plays a crucial role in our analysis, has an impact on the actual performance of the algorithm.

****

Mixture Regression for Covariate Shift

Masashi Sugiyama, Amos J. Storkey

In supervised learning there is a typical presumption that the training and test points are taken from the same distribution. In practice this assumption is commonly violated. The situations where the training and test data are from different distributions is called covariate shift. Recent work has examined techniques for dealing with covariate shift in terms of minimisation of generalisation error. As yet the literature lacks a Bayesian generative perspective on this problem. This paper tackles this issue for regression models. Recent work on covariate shift can be understood in terms of mixture regression. Using this view, we obtain a general approach to regression under covariate shift, which reproduces previous work as a special case. The main advantages of this new formulation over previous models for covariate shift are that we no longer need to presume the test and training densities are known, the regression and density estimation are combined into a single procedure, and previous methods are reproduced as special cases of this procedure, shedding light on the implicit assumptions the methods are making.

****

implicit Online Learning with Kernels

Li Cheng, Dale Schuurmans, Shaojun Wang, Terry Caelli, S.v.n. Vishwanathan

We present two new algorithms for online learning in reproducing kernel Hilbert spaces. Our first algorithm, ILK (implicit online learning with kernels), employs a new, implicit update technique that can be applied to a wide variety of convex loss functions. We then introduce a bounded memory version, SILK (sparse ILK), that maintains a compact representation of the predictor without compromising solution quality, even in non-stationary environments. We prove loss bounds and analyze the convergence rate of both. Experimental evidence shows that our proposed algorithms outperform current methods on synthetic and real data.

****

PG-means: learning the number of clusters in data

Yu Feng, Greg Hamerly

We present a novel algorithm called PG-means which is able to learn the number of clusters in a classical Gaussian mixture model. Our method is robust and efficient; it uses statistical hypothesis tests on one-dimensional projections of the data and model to determine if the examples are well represented by the model. In so doing, we are applying a statistical test for the entire model at once, not just on a per-cluster basis. We show that our method works well in difficult cases such as non-Gaussian data, overlapping clusters, eccentric clusters, high dimension, and many true clusters. Further, our new method provides a much more s table estimate of the number of clusters than existing methods.

****

Conditional Random Sampling: A Sketch-based Sampling Technique for Sparse Data

Ping Li, Kenneth Church, Trevor Hastie

We1 develop Conditional Random Sampling (CRS), a technique particularly suit- able for sparse data. In large-scale applications, the data are often highly sparse. CRS combines sketching and sampling in that it converts sketches of the data into conditional random samples online in the estimation stage, with the sample size determined retrospectively. This paper focuses on approximating pairwise l2 and l1 distances and comparing CRS with random projections. For boolean (0/1) data, CRS is provably better than random projections. We show using real-world data that CRS often outperforms random projections. This technique can be applied in learning, data mining, information retrieval, and database query optimizations.

****

Multi-Robot Negotiation: Approximating the Set of Subgame Perfect Equilibria in General-Sum Stochastic Games

Chris Murray, Geoffrey J. Gordon

In real-world planning problems, we must reason not only about our own goals, but about the goals of other agents with which we may interact. Often these agents' goals are neither completely aligned with our own nor directly opposed to them. Instead there are opportunities for cooperation: by joining forces, the agents can all achieve higher utility than they could separately. But, in order to cooperate, the agents must negotiate a mutually acceptable plan from among the many possible ones, and each agent must trust that the others will follow their parts of the deal. Research in multi-agent planning has often avoided the problem of making sure that all agents have an incentive to follow a proposed joint plan. On the other hand, while game theoretic algorithms handle incentives correctly, they often don't scale to large planning problems. In this paper we attempt to bridge the gap between these two lines of research: we present an efficient game-theoretic approximate planning algorithm, along with a negotiation protocol which encourages agents to compute and agree on joint plans that are fair and optimal in a sense defined below. We demonstrate our algorithm and protocol on two simple robotic planning problems.1

****

Optimal Single-Class Classification Strategies

Ran El-Yaniv, Mordechai Nisenson

We consider single-class classification (SCC) as a two-person game between the learner and an adversary. In this game the target distribution is completely known to the learner and the learner's goal is to construct a classifier capable of guaranteeing a given tolerance for the false-positive error while minimizing the false negative error. We identify both "hard" and "soft" optimal classification strategies for different types of games and demonstrate that soft classification can provide a significant advantage. Our optimal strategies and bounds provide worst-case lower bounds for standard, finite-sample SCC and also motivate new approaches to solving SCC.

****

Learning to Rank with Nonsmooth Cost Functions

Christopher Burges, Robert Ragno, Quoc Le

The quality measures used in information retrieval are particularly difficult to op- timize directly, since they depend on the model scores only through the sorted order of the documents returned for a given query. Thus, the derivatives of the cost with respect to the model parameters are either zero, or are undefined. In this paper, we propose a class of simple, flexible algorithms, called LambdaRank, which avoids these difficulties by working with implicit cost functions. We de- scribe LambdaRank using neural network models, although the idea applies to any differentiable function class. We give necessary and sufficient conditions for the resulting implicit cost function to be convex, and we show that the general method has a simple mechanical interpretation. We demonstrate significantly im- proved accuracy, over a state-of-the-art ranking algorithm, on several datasets.

We also show that LambdaRank provides a method for significantly speeding up the training phase of that ranking algorithm. Although this paper is directed towards ranking, the proposed method can be extended to any non-smooth and multivariate cost functions.
****
Analysis of Representations for Domain Adaptation
Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira
Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. In many situations, though, we have labeled training data for a source domain, and we wish to learn a classifier which performs well on a target domain with a different distribution. Under what conditions can we adapt a classifier trained on the source domain for use in the target domain? Intuitively, a good feature representation is a crucial factor in the success of domain adaptation. We formalize this intuition theoretically with a generalization bound for domain adaption. Our theory illustrates the tradeoffs inherent in designing a representation for domain adaptation and gives a new justification for a recently proposed model. It also points toward a promising new model for domain adaptation: one which explicitly minimizes the difference between the source and target domains, while at the same time maximizing the margin of the training set.
****
Neurophysiological Evidence of Cooperative Mechanisms for Stereo Computation
Jason Samonds, Brian Potetz, Tai Lee
Although there has been substantial progress in understanding the neurophysiological mechanisms of stereopsis, how neurons interact in a network during stereo computation remains unclear. Computational models on stereopsis suggest local competition and long-range cooperation are important for resolving ambiguity during stereo matching. To test these predictions, we simultaneously recorded from multiple neurons in V1 of awake, behaving macaques while presenting surfaces of different depths rendered in dynamic random dot stereograms. We found that the interaction between pairs of neurons was a function of similarity in receptive fields, as well as of the input stimulus. Neurons coding the same depth experienced common inhibition early in their responses for stimuli presented at their non-preferred disparities. They experienced mutual facilitation later in their responses for stimulation at their preferred disparity. These findings are consistent with a local competition mechanism that first removes gross mismatches, and a global cooperative mechanism that further refines depth estimates.
****
Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation
Gavin Cawley, Nicola Talbot, Mark Girolami
Multinomial logistic regression provides the standard penalised maximum- likelihood solution to multi-class pattern recognition problems. More recently, the development of sparse multinomial logistic regression models has found ap- plication in text processing and microarray classification, where explicit identifi- cation of the most informative features is of value. In this paper, we propose a sparse multinomial logistic regression method, in which the sparsity arises from the use of a Laplace prior, but where the usual regularisation parameter is inte- grated out analytically. Evaluation over a range of benchmark datasets reveals this approach results in similar generalisation performance to that obtained using cross-validation, but at greatly reduced computational expense.
****
Gaussian and Wishart Hyperkernels
Risi Kondor, Tony Jebara
We propose a new method for constructing hyperkenels and define two promising special cases that can be computed in closed form. These we call the Gaussian and Wishart hyperkernels. The former is especially attractive in that it has an interpretable regularization scheme reminiscent of that of the Gaussian RBF kernel. We discuss how kernel learning can be used not just for improving the performanc

e of classification and regression methods, but also as a stand-alone algorithm for dimensionality reduction and relational or metric learning.
****
A Complexity-Distortion Approach to Joint Pattern Alignment
Andrea Vedaldi, Stefano Soatto
Image Congealing (IC) is a non-parametric method for the joint alignment of a col- lection of images affected by systematic and unwanted deformations. The method attempts to undo the deformations by minimizing a measure of complexity of the image ensemble, such as the averaged per-pixel entropy. This enables alignment without an explicit model of the aligned dataset as required by other methods (e.g. transformed component analysis). While IC is simple and general, it may intro- duce degenerate solutions when the transformations allow minimizing the com- plexity of the data by collapsing them to a constant. Such solutions need to be explicitly removed by regularization. In this paper we propose an alternative formulation which solves this regulariza- tion issue on a more principled ground. We make the simple observation that alignment should simplify the data while preserving the useful information car- ried by them. Therefore we trade off fidelity and complexity of the aligned en- semble rather than minimizing the complexity alone. This eliminates the need for an explicit regularization of the transformations, and has a number of other useful properties such as noise suppression. We show the modeling and computa- tional benefits of the approach to the some of the problems on which IC has been demonstrated.
****
AdaBoost is Consistent
Peter Bartlett, Mikhail Traskin
The risk, or probability of error, of the classifier produced by the AdaBoost algorithm is investigated. In particular, we consider the stopping strategy to be used in AdaBoost to achieve universal consistency. We show that provided AdaBoost is stopped after n iterations--for sample size n and < 1--the sequence of risks of the classifiers it produces approaches the Bayes risk if Bayes risk L > 0.
****
Cross-Validation Optimization for Large Scale Hierarchical Classification Kernel Methods
Matthias Seeger
We propose a highly efficient framework for kernel multi-class models with a large and structured set of classes. Kernel parameters are learned automatically by maximizing the cross-validation log likelihood, and predictive probabilities are estimated. We demonstrate our approach on large scale text classification tasks with hierarchical class structure, achieving state-of-the-art results in an order of magnitude less time than previous work.
****
Effects of Stress and Genotype on Meta-parameter Dynamics in Reinforcement Learning
Gediminas Lukšys, Jérémie Knüsel, Denis Sheynikhovich, Carmen Sandi, Wulfram Gerstner
Stress and genetic background regulate different aspects of behavioral learning through the action of stress hormones and neuromodulators. In reinforcement learning (RL) models, meta-parameters such as learning rate, future reward dis- count factor, and exploitation-exploration factor, control learning dynamics and performance. They are hypothesized to be related to neuromodulatory levels in the brain. We found that many aspects of animal learning and performance can be described by simple RL models using dynamic control of the meta-parameters. To study the effects of stress and genotype, we carried out 5-hole-box light condition- ing and Morris water maze experiments with C57BL/6 and DBA/2 mouse strains. The animals were exposed to different kinds of stress to evaluate its effects on immediate performance as well as on long-term memory. Then, we used RL mod- els to simulate their behavior. For each experimental session, we estimated a set of model meta-parameters that produced the best fit between the model and the animal performance. The dynamics of several estimated meta-parameters were qualitatively similar for the two simulated experiments, and with statistically sig- nificant d

ifferences between different genetic strains and stress conditions.

****

A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems

David Barber, Bertrand Mesot

We introduce a method for approximate smoothed inference in a class of switching linear dynamical systems, based on a novel form of Gaussian Sum smoother. This class includes the switching Kalman Filter and the more general case of switch transitions dependent on the continuous latent state. The method improves on the standard Kim smoothing approach by dispensing with one of the key approximations, thus making fuller use of the available future information. Whilst the only central assumption required is projection to a mixture of Gaussians, we show that an additional conditional independence assumption results in a simpler but stable and accurate alternative. Unlike the alternative unstable Expectation Propagation procedure, our method consists only of a single forward and backward pass and is reminiscent of the standard smoothing `correction' recursions in the simpler linear dynamical system. The algorithm performs well on both toy experiments and in a large scale application to noise robust speech recognition.

****

Information Bottleneck Optimization and Independent Component Extraction with Spiking Neurons

Stefan Klampfl, Wolfgang Maass, Robert Legenstein

The extraction of statistically independent components from high-dimensional multi-sensory input streams is assumed to be an essential component of sensory processing in the brain. Such independent component analysis (or blind source separation) could provide a less redundant representation of information about the external world. Another powerful processing strategy is to extract preferentially those components from high-dimensional input streams that are related to other information sources, such as internal predictions or proprioceptive feedback. This strategy allows the optimization of internal representation according to the infor- mation bottleneck method. However, concrete learning rules that implement these general unsupervised learning principles for spiking neurons are still missing. We show how both information bottleneck optimization and the extraction of inde- pendent components can in principle be implemented with stochastically spiking neurons with refractoriness. The new learning rule that achieves this is derived from abstract information optimization principles.

****

A Theory of Retinal Population Coding

Eizaburo Doi, Michael Lewicki

Efficient coding models predict that the optimal code for natural images is a population of oriented Gabor receptive fields. These results match response properties of neurons in primary visual cortex, but not those in the retina. Does the retina use an optimal code, and if so, what is it optimized for? Previous theories of retinal coding have assumed that the goal is to encode the maximal amount of information about the sensory signal. However, the image sampled by retinal photoreceptors is degraded both by the optics of the eye and by the photoreceptor noise. Therefore, de-blurring and de-noising of the retinal signal should be important aspects of retinal coding. Furthermore, the ideal retinal code should be robust to neural noise and make optimal use of all available neurons. Here we present a theoretical framework to derive codes that simultaneously satisfy all of these desiderata. When optimized for natural images, the model yields filters that show strong similarities to retinal ganglion cell (RGC) receptive fields. Importantly, the characteristics of receptive fields vary with retinal eccentricities where the optical blur and the number of RGCs are significantly different. The proposed model provides a unified account of retinal coding, and more generally, it may be viewed as an extension of the Wiener filter with an arbitrary number of noisy units.

****

An Approach to Bounded Rationality

Eli Ben-sasson, Ehud Kalai, Adam Kalai

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

****

Fundamental Limitations of Spectral Clustering

Boaz Nadler, Meirav Galun

Spectral clustering methods are common graph-based approaches to clustering of data. Spectral clustering algorithms typically start from local information encoded in a weighted graph on the data and cluster according to the global eigenvectors of the corresponding (normalized) similarity matrix. One contribution of this paper is to present fundamental limitations of this general local to global approach. We show that based only on local information, the normalized cut functional is not a suitable measure for the quality of clustering. Further, even with a suitable similarity measure, we show that the first few eigenvectors of such adjacency matrices cannot successfully cluster datasets that contain structures at different scales of size and density. Based on these findings, a second contribution of this paper is a novel diffusion based measure to evaluate the coherence of individual clusters. Our measure can be used in conjunction with any bottom-up graph-based clustering method, it is scale-free and can determine coherent clusters at all scales. We present both synthetic examples and real image segmentation problems where various spectral clustering algorithms fail. In contrast, using this coherence measure finds the expected clusters at all scales. Keywords: Clustering, kernels, learning theory.

****

Generalized Maximum Margin Clustering and Unsupervised Kernel Learning

Hamed Valizadegan, Rong Jin

Maximum margin clustering was proposed lately and has shown promising performance in recent studies [1, 2]. It extends the theory of support vector machine to unsupervised learning. Despite its good performance, there are three ma jor problems with maximum margin clustering that question its efficiency for real-world applications. First, it is computationally expensive and difficult to scale to large-scale datasets because the number of parameters in maximum margin clustering is quadratic in the number of examples. Second, it requires data preprocessing to ensure that any clustering boundary will pass through the origins, which makes it unsuitable for clustering unbalanced dataset. Third, it is sensitive to the choice of kernel functions, and requires external procedure to determine the appropriate values for the parameters of kernel functions. In this paper, we propose "generalized maximum margin clustering" framework that addresses the above three problems simultaneously. The new framework generalizes the maximum margin clustering algorithm by allowing any clustering boundaries including those not passing through the origins. It significantly improves the computational efficiency by reducing the number of parameters. Furthermore, the new framework is able to automatically determine the appropriate kernel matrix without any labeled data. Finally, we show a formal connection between maximum margin clustering and spectral clustering. We demonstrate the efficiency of the generalized maximum margin clustering algorithm using both synthetic datasets and real datasets from the UCI repository.

****

Clustering Under Prior Knowledge with Application to Image Segmentation

Dong Cheng, Vittorio Murino, Mário Figueiredo

This paper proposes a new approach to model-based clustering under prior knowl-edge. The proposed formulation can be interpreted from two different angles: as penalized logistic regression, where the class labels are only indirectly observed (via the probability density of each class); as finite mixture learning under a group- ing prior. To estimate the parameters of the proposed model, we derive a (gener- alized) EM algorithm with a closed-form E-step, in contrast with other recent approaches to semi-supervised probabilistic clustering which require Gibbs sam- pling or suboptimal shortcuts. We show that our approach is ideally suited for image segmentation: it avoids the combinatorial nature Markov random field pri- ors, and opens the door to more sophisticated spatial priors (e.g., wavelet-based) in a simple and computationally efficient way. Finally, we extend our formulation to work in unsupervised, semi-supervised, or discriminative modes.

****

Inferring Network Structure from Co-Occurrences
Michael Rabbat, Mário Figueiredo, Robert Nowak

We consider the problem of inferring the structure of a network from cooccurrence data: observations that indicate which nodes occur in a signaling pathway but do not directly reveal node order within the pathway. This problem is motivated by network inference problems arising in computational biology and communication systems, in which it is difficult or impossible to obtain precise time ordering information. Without order information, every permutation of the activated nodes leads to a different feasible solution, resulting in combinatorial explosion of the feasible set. However, physical principles underlying most networked systems suggest that not all feasible solutions are equally likely. Intuitively, nodes that co-occur more frequently are probably more closely connected. Building on this intuition, we model path co-occurrences as randomly shuffled samples of a random walk on the network. We derive a computationally efficient network inference algorithm and, via novel concentration inequalities for importance sampling estimators, prove that a polynomial complexity Monte Carlo version of the algorithm converges with high probability.

****

Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning
Peter Auer, Ronald Ortner

We present a learning algorithm for undiscounted reinforcement learning. Our interest lies in bounds for the algorithm's online performance after some finite number of steps. In the spirit of similar methods already successfully applied for the exploration-exploitation tradeoff in multi-armed bandit problems, we use upper confidence bounds to show that our UCRL algorithm achieves logarithmic online regret in the number of steps taken with respect to an optimal policy.

****

Fast Iterative Kernel PCA
Nicol Schraudolph, Simon Günter, S.v.n. Vishwanathan

We introduce two methods to improve convergence of the Kernel Hebbian Algorithm (KHA) for iterative kernel PCA. KHA has a scalar gain parameter which is either held constant or decreased as 1/t, leading to slow convergence. Our KHA/et algorithm accelerates KHA by incorporating the reciprocal of the current estimated eigenvalues as a gain vector. We then derive and apply Stochastic MetaDescent (SMD) to KHA/et; this further speeds convergence by performing gain adaptation in RKHS. Experimental results for kernel PCA and spectral clustering of USPS digits as well as motion capture and image de-noising problems confirm that our methods converge substantially faster than conventional KHA.

****

Uncertainty, phase and oscillatory hippocampal recall
Máté Lengyel, Peter Dayan

Many neural areas, notably, the hippocampus, show structured, dynamical, population behavior such as coordinated oscillations. It has long been observed that such oscillations provide a substrate for representing analog information in the firing phases of neurons relative to the underlying population rhythm. However, it has become increasingly clear that it is essential for neural populations to represent uncertainty about the information they capture, and the substantial recent work on neural codes for uncertainty has omitted any analysis of oscillatory systems. Here, we observe that, since neurons in an oscillatory network need not only fire once in each cycle (or even at all), uncertainty about the analog quantities each neuron represents by its firing phase might naturally be reported through the degree of concentration of the spikes that it fires. We apply this theory to memory in a model of oscillatory associative recall in hippocampal area CA3. Although it is not well treated in the literature, representing and manipulating uncertainty is fundamental to competent memory; our theory enables us to view CA3 as an effective uncertainty-aware, retrieval system.

****

Speakers optimize information density through syntactic reduction
T. Jaeger, Roger Levy

If language users are rational, they might choose to structure their utterances so as to optimize communicative properties. In particular, information-theoretic and psycholinguistic considerations suggest that this may include maximizing the uniformity of information density in an utterance. We investigate this possibility in the context of syntactic reduction, where the speaker has the option of either marking a higher-order unit (a phrase) with an extra word, or leaving it unmarked. We demonstrate that speakers are more likely to reduce less information-dense phrases. In a second step, we combine a stochastic model of structured utterance production with a logistic-regression model of syntactic reduction to study which types of cues speakers employ when estimating the predictability of upcoming elements. We demonstrate that the trend toward predictability-sensitive syntactic reduction (Jaeger, 2006) is robust in the face of a wide variety of control variables, and present evidence that speakers use both surface and structural cues for predictability estimation.

****

Learning Structural Equation Models for fMRI
Enrico Simonotto, Heather Whalley, Stephen Lawrie, Lawrence Murray, David Mcgonigle, Amos J. Storkey

David McGonigle

****

Simplifying Mixture Models through Function Approximation
Kai Zhang, James Kwok

Finite mixture model is a powerful tool in many statistical learning problems. In this paper, we propose a general, structure-preserving approach to reduce its model complexity, which can bring significant computational benefits in many applications. The basic idea is to group the original mixture components into compact clusters, and then minimize an upper bound on the approximation error between the original and simplified models. By adopting the L2 norm as the dis- tance measure between mixture models, we can derive closed-form solutions that are more robust and reliable than using the KL-based distance measure. Moreover, the complexity of our algorithm is only linear in the sample size and dimensional- ity. Experiments on density estimation and clustering-based image segmentation demonstrate its outstanding performance in terms of both speed and accuracy.

****

Sparse Representation for Signal Classification
Ke Huang, Selin Aviyente

In this paper, application of sparse representation (factorization) of signals over an overcomplete basis (dictionary) for signal classification is discussed. Search- ing for the sparse representation of a signal over an overcomplete dictionary is achieved by optimizing an objective function that includes two terms: one that measures the signal reconstruction error and another that measures the sparsity. This objective function works well in applications where signals need to be recon- structed, like coding and denoising. On the other hand, discriminative methods, such as linear discriminative analysis (LDA), are better suited for classification tasks. However, discriminative methods are usually sensitive to corruption in sig- nals due to lacking crucial properties for signal reconstruction. In this paper, we present a theoretical framework for signal classification with sparse representa- tion. The approach combines the discrimination power of the discriminative meth- ods with the reconstruction property and the sparsity of the sparse representation that enables one to deal with signal corruptions: noise, missing data and outliers. The proposed approach is therefore capable of robust classification with a sparse representation of signals. The theoretical results are demonstrated with signal classification tasks, showing that the proposed approach outperforms the standard discriminative methods and the standard sparse representation in the case of cor- rupted signals.

****

Similarity by Composition

Oren Boiman, Michal Irani

We propose a new approach for measuring similarity between two signals, which is applicable to many machine learning tasks, and to many signal types. We say that a signal S1 is “similar” to a signal S2 if it is “easy” to compose S1 from few large contiguous chunks of S2. Obviously, if we use small enough pieces, then any signal can be composed of any other. Therefore, the larger those pieces are, the more similar S1 is to S2. This induces a local similarity score at every point in the signal, based on the size of its supported surrounding region. These local scores can in turn be accumulated in a principled information-theoretic way into a global similarity score of the entire S1 to S2. “Similarity by Composition” can be applied between pairs of signals, between groups of signals, and also between dif- ferent portions of the same signal. It can therefore be employed in a wide variety of machine learning problems (clustering, classification, retrieval, segmentation, attention, saliency, labelling, etc.), and can be applied to a wide range of signal types (images, video, audio, biological data, etc.) We show a few such examples.

****

An Information Theoretic Framework for Eukaryotic Gradient Sensing

Joseph Kimmel, Richard Salter, Peter Thomas

Chemical reaction networks by which individual cells gather and process information about their chemical environments have been dubbed "signal transduction" networks. Despite this suggestive terminology, there have been few attempts to analyze chemical signaling systems with the quantitative tools of information theory. Gradient sensing in the social amoeba Dictyostelium discoideum is a well characterized signal transduction system in which a cell estimates the direction of a source of diffusing chemoattractant molecules based on the spatiotemporal sequence of ligand-receptor binding events at the cell membrane. Using Monte Carlo techniques (MCell) we construct a simulation in which a collection of individual ligand particles undergoing Brownian diffusion in a three-dimensional volume interact with receptors on the surface of a static amoeboid cell. Adapting a method for estimation of spike train entropies described by Victor (originally due to Kozachenko and Leonenko), we estimate lower bounds on the mutual information between the transmitted signal (direction of ligand source) and the received signal (spatiotemporal pattern of receptor binding/unbinding events). Hence we provide a quantitative framework for addressing the question: how much could the cell know, and when could it know it? We show that the time course of the mutual information between the cell's surface receptors and the (unknown) gradient direction is consistent with experimentally measured cellular response times. We find that the acquisition of directional information depends strongly on the time constant at which the intracellular response is filtered.

****

Prediction on a Graph with a Perceptron

Mark Herbster, Massimiliano Pontil

We study the problem of online prediction of a noisy labeling of a graph with the perceptron. We address both label noise and concept noise. Graph learning is framed as an instance of prediction on a finite set. To treat label noise we show that the hinge loss bounds derived by Gentile [1] for online perceptron learning can be transformed to relative mistake bounds with an optimal leading constant when applied to prediction on a finite set. These bounds depend crucially on the norm of the learned concept. Often the norm of a concept can vary dramatically with only small perturbations in a labeling. We analyze a simple transformation that stabilizes the norm under perturbations. We derive an upper bound that depends only on natural properties of the graph the graph diameter and the cut size of a partitioning of the graph which are only indirectly dependent on the size of the graph. The impossibility of such bounds for the graph geodesic nearest neighbors algorithm will be demonstrated.

****

An Efficient Method for Gradient-Based Adaptation of Hyperparameters in SVM Models

S. Keerthi, Vikas Sindhwani, Olivier Chapelle

We consider the task of tuning hyperparameters in SVM models based on minimizing a smooth performance validation function, e.g., smoothed k-fold crossvalidation error, using non-linear optimization techniques. The key computation in this approach is that of the gradient of the validation function with respect to hyperparameters. We show that for large-scale problems involving a wide choice of kernel-based models and validation functions, this computation can be very efficiently done; often within just a fraction of the training time. Empirical results show that a near-optimal set of hyperparameters can be identified by our approach with very few training rounds and gradient computations. .

****

Analysis of Empirical Bayesian Methods for Neuroelectromagnetic Source Localization

Rey Ramírez, Jason Palmer, Scott Makeig, Bhaskar Rao, David Wipf

The ill-posed nature of the MEG/EEG source localization problem requires the incorporation of prior assumptions when choosing an appropriate solution out of an infinite set of candidates. Bayesian methods are useful in this capacity because they allow these assumptions to be explicitly quantified. Recently, a number of empirical Bayesian approaches have been proposed that attempt a form of model selection by using the data to guide the search for an appropriate prior. While seemingly quite different in many respects, we apply a unifying framework based on automatic relevance determination (ARD) that elucidates various attributes of these methods and suggests directions for improvement. We also derive theoretical properties of this methodology related to convergence, local minima, and localization bias and explore connections with established algorithms.

****

Efficient Methods for Privacy Preserving Face Detection

Shai Avidan, Moshe Butman

Bob offers a face-detection web service where clients can submit their images for analysis. Alice would very much like to use the service, but is reluctant to reveal the content of her images to Bob. Bob, for his part, is reluctant to release his face detector, as he spent a lot of time, energy and money constructing it. Secure Multi- Party computations use cryptographic tools to solve this problem without leaking any information. Unfortunately, these methods are slow to compute and we intro- duce a couple of machine learning techniques that allow the parties to solve the problem while leaking a controlled amount of information. The first method is an information-bottleneck variant of AdaBoost that lets Bob find a subset of features that are enough for classifying an image patch, but not enough to actually recon- struct it. The second machine learning technique is active learning that allows Alice to construct an online classifier, based on a small number of calls to Bob’s face detector. She can then use her online classifier as a fast rejector before using a cryptographically secure classifier on the remaining image patches.

****

Balanced Graph Matching

Timothee Cour, Praveen Srinivasan, Jianbo Shi

Graph matching is a fundamental problem in Computer Vision and Machine Learning. We present two contributions. First, we give a new spectral relaxation technique for approximate solutions to matching problems, that naturally incorporates one-to-one or one-to-many constraints within the relaxation scheme. The second is a normalization procedure for existing graph matching scoring functions that can dramatically improve the matching accuracy. It is based on a reinterpretation of the graph matching compatibility matrix as a bipartite graph on edges for which we seek a bistochastic normalization. We evaluate our two contributions on a comprehensive test set of random graph matching problems, as well as on image correspondence problem. Our normalization procedure can be used to improve the performance of many existing graph matching algorithms, including spectral matching, graduated assignment and semidefinite programming.

****

Fast Discriminative Visual Codebooks using Randomized Clustering Forests

Frank Moosmann, Bill Triggs, Frederic Jurie

Some of the most effective recent methods for content-based image classification work by extracting dense or sparse local image descriptors, quantizing them according to a coding rule such as k-means vector quantization, accumulating histograms of the resulting "visual word" codes over the image, and classifying these with a conventional classifier such as an SVM. Large numbers of descriptors and large codebooks are needed for good results and this becomes slow using k-means. We introduce Extremely Randomized Clustering Forests ensembles of randomly created clustering trees and show that these provide more accurate results, much faster training and testing and good resistance to background clutter in several state-of-the-art image classification tasks.

****

Learning Motion Style Synthesis from Perceptual Observations

Lorenzo Torresani, Peggy Hackney, Christoph Bregler

This paper presents an algorithm for synthesis of human motion in specified styles. We use a theory of movement observation (Laban Movement Analysis) to describe movement styles as points in a multi-dimensional perceptual space. We cast the task of learning to synthesize desired movement styles as a regression problem: sequences generated via space-time interpolation of motion capture data are used to learn a nonlinear mapping between animation parameters and movement styles in perceptual space. We demonstrate that the learned model can apply a variety of motion styles to pre-recorded motion sequences and it can extrapolate styles not originally included in the training data.

****

Linearly-solvable Markov decision problems

Emanuel Todorov

We introduce a class of MPDs which greatly simplify Reinforcement Learning. They have discrete state spaces and continuous control spaces. The controls have the effect of rescaling the transition probabilities of an underlying Markov chain. A control cost penalizing KL divergence between controlled and uncontrolled transition probabilities makes the minimization problem convex, and allows analytical computation of the optimal controls given the optimal value function. An exponential transformation of the optimal value function makes the minimized Bellman equation linear. Apart from their theoretical significance, the new MDPs enable efficient approximations to traditional MDPs. Shortest path problems are approximated to arbitrary precision with largest eigenvalue problems, yielding an O (n) algorithm. Accurate approximations to generic MDPs are obtained via continuous embedding reminiscent of LP relaxation in integer programming. Offpolicy learning of the optimal value function is possible without need for stateaction values; the new algorithm (Z-learning) outperforms Q-learning. This work was supported by NSF grant ECS0524761.

****

Learning on Graph with Laplacian Regularization

Rie Ando, Tong Zhang

We consider a general form of transductive learning on graphs with Laplacian regularization, and derive margin-based generalization bounds using appropriate geometric properties of the graph. We use this analysis to obtain a better understanding of the role of normalization of the graph Laplacian matrix as well as the effect of dimension reduction. The results suggest a limitation of the standard degree-based normalization. We propose a remedy from our analysis and demonstrate empirically that the remedy leads to improved classification performance.

****

An Oracle Inequality for Clipped Regularized Risk Minimizers

Ingo Steinwart, Don Hush, Clint Scovel

We establish a general oracle inequality for clipped approximate minimizers of regularized empirical risks and apply this inequality to support vector machine (SVM) type algorithms. We then show that for SVMs using Gaussian RBF kernels for classification this oracle inequality leads to learning rates that are faster than the ones established in [9]. Finally, we use our oracle inequality to show that a simple parameter selection approach based on a validation set can yield the same fast learning rates without knowing the noise exponents which were require d to be known a-priori in [9].

****

Clustering appearance and shape by learning jigsaws

Anitha Kannan, John Winn, Carsten Rother

Patch-based appearance models are used in a wide range of computer vision ap- plications. To learn such models it has previously been necessary to specify a suitable set of patch sizes and shapes by hand. In the jigsaw model presented here, the shape, size and appearance of patches are learned automatically from the repeated structures in a set of training images. By learning such irregularly shaped ‘jigsaw pieces’, we are able to discover both the shape and the appearance of object parts without supervision. When applied to face images, for example, the learned jigsaw pieces are surprisingly strongly associated with face parts of different shapes and scales such as eyes, noses, eyebrows and cheeks, to name a few. We conclude that learning the shape of the patch not only improves the accuracy of appearance-based part detection but also allows for shape-based part detection. This enables parts of similar appearance but different shapes to be distinguished; for example, while foreheads and cheeks are both skin colored, they have markedly different shapes.

****

Large Margin Component Analysis

Lorenzo Torresani, Kuang-chih Lee

Metric learning has been shown to significantly improve the accuracy of k-nearest neighbor (kNN) classification. In problems involving thousands of features, distance learning algorithms cannot be used due to overfitting and high computa- tional complexity. In such cases, previous work has relied on a two-step solution: first apply dimensionality reduction methods to the data, and then learn a metric in the resulting low-dimensional subspace. In this paper we show that better classification performance can be achieved by unifying the objectives of dimensionality reduction and metric learning. We propose a method that solves for the low-dimensional projection of the inputs, which minimizes a metric objective aimed at separating points in different classes by a large margin. This projection is defined by a significantly smaller number of parameters than metrics learned in input space, and thus our optimization reduces the risks of overfitting. Theory and results are presented for both a linear as well as a kernelized version of the algorithm. Overall, we achieve classification rates similar, and in several cases superior, to those of support vector machines.

****

Kernels on Structured Objects Through Nested Histograms

Marco Cuturi, Kenji Fukumizu

We propose a family of kernels for structured objects which is based on the bag-ofcomponents paradigm. However, rather than decomposing each complex object into the single histogram of its components, we use for each object a family of nested histograms, where each histogram in this hierarchy describes the object seen from an increasingly granular perspective. We use this hierarchy of histograms to define elementary kernels which can detect coarse and fine similarities between the objects. We compute through an efficient averaging trick a mixture of such specific kernels, to propose a final kernel value which weights efficiently local and global matches. We propose experimental results on an image retrieval experiment which show that this mixture is an effective template procedure to be used with kernels on histograms.

****

Learning with Hypergraphs: Clustering, Classification, and Embedding

Dengyong Zhou, Jiayuan Huang, Bernhard Schölkopf

We usually endow the investigated objects with pairwise relationships, which can be illustrated as graphs. In many real-world problems, however, relationships among the objects of our interest are more complex than pair- wise. Naively squeezing the complex relationships into pairwise ones will inevitably lead to loss of information which can be expected valuable for our learning tasks however. Therefore we consider using hypergraphs in- stead to completely represent complex relationships among the objects of our interest, and thus the problem of learning with hypergraphs arises. Our main contribution in this paper is to generalize the powerful methodology of spectral clustering which originally operates on undirected graphs to hy- pergraphs, and further develop algorithms for hypergraph embedding and transductive classification on the basis of the spectral hypergraph cluster- ing approach. Our experiments on a number of benchmarks showed the advantages of hypergraphs over usual graphs.

****

Learning Dense 3D Correspondence

Florian Steinke, Volker Blanz, Bernhard Schölkopf

Establishing correspondence between distinct objects is an important and nontrivial task: correctness of the correspondence hinges on properties which are difficult to capture in an a priori criterion. While previous work has used a priori criteria which in some cases led to very good results, the present paper explores whether it is possible to learn a combination of features that, for a given training set of aligned human heads, characterizes the notion of correct correspondence. By optimizing this criterion, we are then able to compute correspondence and morphs for novel heads.

****

Fast Computation of Graph Kernels

Karsten Borgwardt, Nicol Schraudolph, S.v.n. Vishwanathan

Using extensions of linear algebra concepts to Reproducing Kernel Hilbert Spaces (RKHS), we define a unifying framework for random walk kernels on graphs. Re- duction to a Sylvester equation allows us to compute many of these kernels in O(n3) worst-case time. This includes kernels whose previous worst-case time complexity was O(n6), such as the geometric kernels of G¨artner et al. [1] and the marginal graph kernels of Kashima et al. [2]. Our algebra in RKHS allow us to exploit sparsity in directed and undirected graphs more effectively than previ- ous methods, yielding sub-cubic computational complexity when combined with conjugate gradient solvers or fixed-point iterations. Experiments on graphs from bioinformatics and other application domains show that our algorithms are often more than 1000 times faster than existing approaches.

****

Multi-dynamic Bayesian Networks

Karim Filali, Jeff A. Bilmes

We present a generalization of dynamic Bayesian networks to concisely describe complex probability distributions such as in problems with multiple interacting variable-length streams of random variables. Our framewor k incorporates recent graphical model constructs to account for existence uncert ainty, value-specific independence, aggregation relationships, and local and global constraints, while still retaining a Bayesian network interpretation and effic ient inference and learning techniques. We introduce one such general technique, which is an extension of Value Elimination, a backtracking search inference algo rithm. Multi-dynamic Bayesian networks are motivated by our work on Statistical Machine Translation (MT). We present results on MT word alignment in support of our claim that MDBNs are a promising framework for the rapid prototyping of new MT systems.

****

Max-margin classification of incomplete data

Gal Chechik, Geremy Heitz, Gal Elidan, Pieter Abbeel, Daphne Koller

We consider the problem of learning classifiers for structurally incomplete data, where some ob jects have a subset of features inherently absent due to complex relationships between the features. The common approach for handling missing features is to begin with a preprocessing phase that completes the missing features, and then use a standard classification procedure. In this paper we show how incomplete data can be classified directly without any completion of the missing features using a max-margin learning framework. We formulate this task using a geometrically-inspired ob jective function, and discuss two optimization approaches: The linearly separable case is written as a set of convex feasibility problems, and the non-separable case has a non-convex ob jective that we optimize iteratively. By avoiding the pre-processing phase in which the data is completed, these approaches offer considerable computational savings. More importantly, we sh ow that by elegantly handling complex patterns of missing values, our approach is both competitive with other methods when the values are missing at random and outperforms them when the missing values have non-trivial structure. We demonstrate our results on two real-world problems: edge prediction in metabolic pathways, and automobile detection in natural images.

****

Scalable Discriminative Learning for Natural Language Parsing and Translation

Joseph Turian, Benjamin Wellington, I. Melamed

Parsing and translating natural languages can be viewed as problems of predicting tree structures. For machine learning approaches to these predictions, the diversity and high dimensionality of the structures involved mandate very large training sets. This paper presents a purely discriminative learning method that scales up well to problems of this size. Its accuracy was at least as good as other comparable methods on a standard parsing task. To our knowledge, it is the first purely discriminative learning algorithm for translation with treestructured models. Unlike other popular methods, this method does not require a great deal of feature engineering a priori, because it performs feature selection over a compound feature space as it learns. Experiments demonstrate the method's versatility, accuracy, and efficiency. Relevant software is freely available at http://nlp.cs.nyu.edu/parser and http://nlp.cs.nyu.edu/GenPar.

****

On the Relation Between Low Density Separation, Spectral Clustering and Graph Cuts

Hariharan Narayanan, Mikhail Belkin, Partha Niyogi

One of the intuitions underlying many graph-based methods for clustering and semi-supervised learning, is that class or cluster boundaries pass through areas of low probability density. In this paper we provide some formal analysis of that notion for a probability distribution. We introduce a notion of weighted boundary volume, which measures the length of the class/cluster boundary weighted by the density of the underlying probability distribution. We show that sizes of the cuts of certain commonly used data adjacency graphs converge to this continuous weighted volume of the boundary.

****

A Kernel Method for the Two-Sample-Problem

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, Alex Smola

We propose two statistical tests to determine if two samples are from different dis- tributions. Our test statistic is in both cases the distance between the means of the two samples mapped into a reproducing kernel Hilbert space (RKHS). The first test is based on a large deviation bound for the test statistic, while the second is based on the asymptotic distribution of this statistic. The test statistic can be com- puted in O(m2) time. We apply our approach to a variety of problems, including attribute matching for databases using the Hungarian marriage method, where our test performs strongly. We also demonstrate excellent performance when compar- ing distributions over graphs, for which no alternative tests currently exist.

****

Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model

Chaitanya Chemudugunta, Padhraic Smyth, Mark Steyvers

Techniques such as probabilistic topic models and latent-semantic indexing have been shown to be broadly useful at automatically extracting the topical or seman - tic content of documents, or more generally for dimension-reduction of sparse count data. These types of models and algorithms can be viewed as generating an abstraction from the words in a document to a lower-dimensional latent variable representation that captures what the document is generally about beyond the spe - cific words it contains. In this paper we propose a new probabilistic model that tempers this approach by representing each document as a combination of (a) a background distribution over common words, (b) a mixture distribution over general topics, and (c) a distribution over words that are treated as being specific to that document. We illustrate how this model can be used for information retr ieval by matching documents both at a general topic level and at a specific word level, providing an advantage over techniques that only match documents at a general level (such as topic models or latent-sematic indexing) or that only match docu- ments at the specific word level (such as TF-IDF).

****

Emergence of conjunctive visual features by quadratic independent component analysis

J.t. Lindgren, Aapo Hyvärinen

In previous studies, quadratic modelling of natural images has resulted in cell models that react strongly to edges and bars. Here we apply quadratic Independent Component Analysis to natural image patches, and show that up to a small approximation error, the estimated components are computing conjunctions of two linear features. These conjunctive features appear to represent not only edges and bars, but also inherently two-dimensional stimuli, such as corners. In addition, we show that for many of the components, the underlying linear features have essentially V1 simple cell receptive field characteristics. Our results indicate that the development of the V2 cells preferring angles and corners may be partly explainable by the principle of unsupervised sparse coding of natural images.

****

Nonlinear physically-based models for decoding motor-cortical population activity

Gregory Shakhnarovich, Sung-phil Kim, Michael Black

Neural motor prostheses (NMPs) require the accurate decoding of motor cortical population activity for the control of an artificial motor system. Previous work on cortical decoding for NMPs has focused on the recovery of hand kinematics. Human NMPs however may require the control of computer cursors or robotic devices with very different physical and dynamical properties. Here we show that the firing rates of cells in the primary motor cortex of non-human primates can be used to control the parameters of an artificial physical system exhibiting realistic dynamics. The model represents 2D hand motion in terms of a point mass connected to a system of idealized springs. The nonlinear spring coefficients are estimated from the firing rates of neurons in the motor cortex. We evaluate linear and a nonlinear decoding algorithms using neural recordings from two monkeys performing two different tasks. We found that the decoded spring coefficients produced accurate hand trajectories compared with state-of-the-art methods for direct decoding of hand kinematics. Furthermore, using a physically-based system produced decoded movements that were more “natural” in that their frequency spectrum more closely matched that of natural hand movements.

****

Conditional mean field

Peter Carbonetto, Nando Freitas

Despite all the attention paid to variational methods based on sum-product message passing (loopy belief propagation, tree-reweighted sum-product), these methods are still bound to inference on a small set of probabilistic models. Mean field approximations have been applied to a broader set of problems, but the solutions are often poor. We propose a new class of conditionally-specified variational approximations based on mean field theory. While not usable on their own, combined with sequential Monte Carlo they produce guaranteed improvements over conventional mean field. Moreover, experiments on a well-studied problem-- inferring the stable configurations of the Ising spin glass--show that the solutions can be significantly better than those obtained using sum-product-based methods.

****

Unsupervised Regression with Applications to Nonlinear System Identification

Ali Rahimi, Ben Recht

We derive a cost functional for estimating the relationship between highdimensional observations and the low-dimensional process that generated them with no input-output examples. Limiting our search to invertible observation functions confers numerous benefits, including a compact representation and no suboptimal local minima. Our approximation algorithms for optimizing this cost functional are fast and give diagnostic bounds on the quality of their solution. Our method can be viewed as a manifold learning algorithm that utilizes a prior on the low-dimensional manifold coordinates. The benefits of taking advantage of such priors in manifold learning and searching for the inverse observation functions in system identification are demonstrated empirically by learning to track moving targets from raw measurements in a sensor network setting and in an RFID tracking experiment.

****

Large Margin Hidden Markov Models for Automatic Speech Recognition

Fei Sha, Lawrence Saul

We study the problem of parameter estimation in continuous density hidden Markov models (CD-HMMs) for automatic speech recognition (ASR). As in support vector machines, we propose a learning algorithm based on the goal of margin maximization. Unlike earlier work on max-margin Markov networks, our approach is specifically geared to the modeling of real-valued observations (such as acoustic feature vectors) using Gaussian mixture models. Unlike previous discriminative frameworks for ASR, such as maximum mutual information and minimum classification error, our framework leads to a convex optimization, without any spurious local minima. The objective function for large margin training of CD-HMMs is defined over a parameter space of positive semidefinite matrices. Its optimization can be performed efficiently with simple gradient-based methods that scale well to large problems. We obtain competitive results for phonetic recognition on the TIMIT speech corpus.

****

No-regret Algorithms for Online Convex Programs

Geoffrey J. Gordon

Online convex programming has recently emerged as a powerful primitive for designing machine learning algorithms. For example, OCP can be used for learning a linear classifier, dynamically rebalancing a binary search tree, finding the shortest path in a graph with unknown edge lengths, solving a structured classification problem, or finding a good strategy in an extensive-form game. Several researchers have designed no-regret algorithms for OCP. But, compared to algorithms for special cases of OCP such as learning from expert advice, these algorithms are not very numerous or flexible. In learning from expert advice, one tool which has proved particularly valuable is the correspondence between no-regret algorithms and convex potential functions: by reasoning about these potential functions, researchers have designed algorithms with a wide variety of useful guarantees such as good performance when the target hypothesis is sparse. Until now, there has been no such recipe for the more general OCP problem, and therefore no ability to tune OCP algorithms to take advantage of properties of the problem or data. In this paper we derive a new class of no-regret learning algorithms for OCP. These Lagrangian Hedging algorithms are based on a general class of potential functions, and are a direct generalization of known learning rules like weighted majority and external-regret matching. In addition to proving regret bounds, we demonstrate our algorithms learning to play one-card poker.

****

A PAC-Bayes Risk Bound for General Loss Functions

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand

We provide a PAC-Bayesian bound for the expected loss of convex combinations of classifiers under a wide class of loss functions (which includes the exponential loss and the logistic loss). Our numerical experiments with Adaboost indicate that the proposed upper bound, computed on the training set, behaves very similarly as the true loss estimated on the testing set.

****

Distributed Inference in Dynamical Systems

Stanislav Funiak, Carlos Guestrin, Rahul Sukthankar, Mark Paskin

We present a robust distributed algorithm for approximate probabilistic inference in dynamical systems, such as sensor networks and teams of mobile robots. Using assumed density filtering, the network nodes maintain a tractable representation of the belief state in a distributed fashion. At each time step, the nodes coordinate to condition this distribution on the observations made throughout the network, and to advance this estimate to the next time step. In addition, we identify a significant challenge for probabilistic inference in dynamical systems: message losses or network partitions can cause nodes to have inconsistent beliefs about the current state of the system. We address this problem by developing distributed algorithms that guarantee that nodes will reach an informative consistent distribution when communication is re-established. We present a suite of experimental results on real-world sensor data for two real sensor network deployments: one with 25 cameras and another with 54 temperature sensors.

****

Modelling transcriptional regulation using Gaussian Processes

Neil Lawrence, Guido Sanguinetti, Magnus Rattray

Modelling the dynamics of transcriptional processes in the cell requires the knowledge of a number of key biological quantities. While some of them are relatively easy to measure, such as mRNA decay rates and mRNA abundance levels, it is still very hard to measure the active concentration levels of the transcription factor proteins that drive the process and the sensitivity of target genes to these concentrations. In this paper we show how these quantities for a given transcription factor can be inferred from gene expression levels of a set of known target genes. We treat the protein concentration as a latent function with a Gaussian process prior, and include the sensitivities, mRNA decay rates and baseline expression levels as hyperparameters. We apply this procedure to a human leukemia dataset, focusing on the tumour repressor p53 and obtaining results in good accordance with recent biological studies.

****

TrueSkill™: A Bayesian Skill Rating System

Ralf Herbrich, Tom Minka, Thore Graepel

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.

****

Learnability and the doubling dimension

Yi Li, Philip Long

metric dimension).

****

Sample Complexity of Policy Search with Known Dynamics

Peter Bartlett, Ambuj Tewari

We consider methods that try to find a good policy for a Markov decision process by choosing one from a given class. The policy is chosen based on its empirical performance in simulations. We are interested in conditions on the complexity of the policy class that ensure the success of such simulation based policy search methods. We show that under bounds on the amount of computation involved in computing policies, transition dynamics and rewards, uniform convergence of empirical estimates to true value functions occurs. Previously, such results were derived by assuming boundedness of pseudodimension and Lipschitz continuity. These assumptions and ours are both stronger than the usual combinatorial complexity measures. We show, via minimax inequalities, that this is essential: boundedness of pseudodimension or fat-shattering dimension alone is not sufficient.

****

Large Scale Hidden Semi-Markov SVMs

Gunnar Rätsch, Sören Sonnenburg

We describe Hidden Semi-Markov Support Vector Machines (SHM SVMs), an extension of HM SVMs to semi-Markov chains. This allows us to predict segmentations of sequences based on segment-based features measuring properties such as the length of the segment. We propose a novel technique to partition the problem into sub-problems. The independently obtained partial solutions can then be recombined in an efficient way, which allows us to solve label sequence learning problems with several thousands of labeled sequences. We have tested our algorithm for predicting gene structures, an important problem in computational biology. Results on a well-known model organism illustrate the great potential of SHM SVMs in computa tional biology.

****

MLLE: Modified Locally Linear Embedding Using Multiple Weights

Zhenyue Zhang, Jing Wang

The locally linear embedding (LLE) is improved by introducing multiple linearly independent local weight vectors for each neighborhood. We characterize the reconstruction weights and show the existence of the linearly independent weight vectors at each neighborhood. The modified locally linear embedding (MLLE) proposed in this paper is much stable. It can retrieve the ideal embedding if MLLE is applied on data points sampled from an isometric manifold. MLLE is also compared with the local tangent space alignment (LTSA). Numerical examples are given that show the improvement and efficiency of MLLE.

****

Hyperparameter Learning for Graph Based Semi-supervised Learning Algorithms

Xinhua Zhang, Wee Lee

Semi-supervised learning algorithms have been successfully applied in many applications with scarce labeled data, by utilizing the unlabeled data. One important category is graph based semi-supervised learning algorithms, for which the performance depends considerably on the quality of the graph, or its hyperparameters. In this paper, we deal with the less explored problem of learning the graphs. We propose a graph learning method for the harmonic energy minimization method; this is done by minimizing the leave-one-out prediction error on labeled data points. We use a gradient based method and designed an efficient algorithm which significantly accelerates the calculation of the gradient by applying the matrix inversion lemma and using careful pre-computation. Experimental results show that the graph learning method is effective in improving the performance of the classification algorithm.

****

Boosting Structured Prediction for Imitation Learning

J. Bagnell, Joel Chestnutt, David Bradley, Nathan Ratliff

The Maximum Margin Planning (MMP) (Ratliff et al., 2006) algorithm solves imitation learning problems by learning linear mappings from features to cost functions in a planning domain. The learned policy is the result of minimum-cost planning using these cost functions. These mappings are chosen so that example policies (or trajectories) given by a teacher appear to be lower cost (with a lossscaled margin) than any other policy for a given planning domain. We provide a novel approach, MMPBOOST , based on the functional gradient descent view of boosting (Mason et al., 1999; Friedman, 1999a) that extends MMP by "boosting" in new features. This approach uses simple binary classification or regression to improve performance of MMP imitation learning, and naturally extends to the class of structured maximum margin prediction problems. (Taskar et al., 2005) Our technique is applied to navigation and planning problems for outdoor mobile robots and robotic legged locomotion.

****

Denoising and Dimension Reduction in Feature Space

Mikio Braun, Klaus-Robert Müller, Joachim Buhmann

We show that the relevant information about a classification problem in feature space is contained up to negligible error in a finite number of leading kernel PCA components if the kernel matches the underlying learning problem. Thus, kernels not only transform data sets such that good generalization can be achieved even by linear discriminant functions, but this transformation is also performed in a manner which makes economic use of feature space dimensions. In the best case, kernels provide efficient implicit representations of the data to perform classification. Practically, we propose an algorithm which enables us to recover the subspace and dimensionality relevant for good classification. Our algorithm can therefore be applied (1) to analyze the interplay of data set and kernel in a geometric fashion, (2) to help in model selection, and to (3) de-noise in feature space in order to yield better classification results.

****

Relational Learning with Gaussian Processes

Wei Chu, Vikas Sindhwani, Zoubin Ghahramani, S. Keerthi

Correlation between instances is often modelled via a kernel function using input attributes of the instances. Relational knowledge can further reveal additional pairwise correlations between variables of interest. In this paper, we develop a class of models which incorporates both reciprocal relational information and in- put attributes using Gaussian process techniques. This approach provides a novel non-parametric Bayesian framework with a data-dependent covariance function for supervised learning tasks. We also apply this framework to semi-supervised learning. Experimental results on several real world data sets verify the usefulness of this algorithm.

****