AISTATS
ARTIFICIAL INTELLIGENCE & STATISTICS CONFERENCE

PASCAL2
Pattern Analysis, Statistical Modelling and
Computational Learning

Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2010

May 13-15, 2010, Chia Laguna, Sardinia, Italy

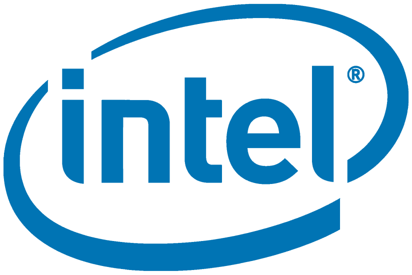
Abstracts of Papers

AISTATS 2010 Sponsors

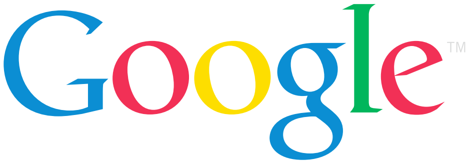
Platinum sponsors

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Silver sponsors

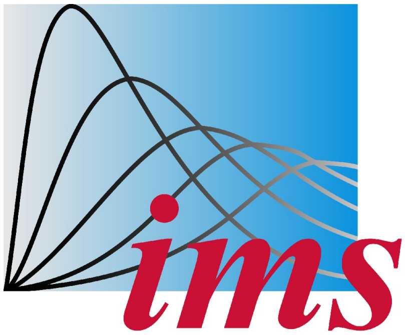


Bronze sponsors



Other Sponsorship

AISTATS is also sponsored by the Institute of Mathematical Statsitics.



Schedule

Wednesday 12 May

16:00 - 19:00	Registration
19:30 - 21:30	Dinner

Thursday 13 May

07:45 - 08:30	Breakfast
08:30 - 08:45	Welcome *Organizers*
	Invited Talk
08:45 - 09:45	Forensic statistics: where are we and where are we going? *Richard Gill*
	Network Models
09:45 - 10:10	Boosted optimization for network classification *T. Hancock and H. Mamitsuka*
10:10 - 10:35	Detecting weak but hierarchically-structured patterns in networks *A. Singh and R. Nowak*
10:35 - 13:00	Coffee Break and Poster Session I
13:00 - 17:00	Lunch
	Statistical Learning Theory
17:00 - 17:25	Risk bounds for transduction and semi-supervised learning relative to data structure *G. Lever*
17:25 - 17:50	Multiclass-multilabel classification with more labels than examples *O. Dekel and O. Shamir*
17:50 - 18:15	Empirical Bernstein boosting *P. Shivaswamy and T. Jebara*
18:15 - 18:45	Tea Break

	Bayesian nonparametrics and causal inference
18:45 - 19:10	Sufficient covariates and linear propensity analysis *H. Guo and P. Dawid*
19:10 - 19:35	Dirichlet process mixtures of generalised linear models *L. Hannah, D. Blei and W. Powell*
19:35 - 20:00	Bayesian Gaussian process latent variable model *M. Titsias and N. Lawrence*
20:00 - 22:00	Conference Banquet

Friday 14 May

07:45 - 08:45	Breakfast
	Invited Talk
08:45 - 09:45	Approximate Bayesian Computation: What, Why and How? *Simon Tavaré*
	Deep Learning
09:45 - 10:10	Factored 3-way restricted Boltzmann machines for modeling natural images *M. Ranzato, A. Krizhevsky and G. Hinton*
10:10 - 10:35	Learning the structure of deep sparse graphical models *R. Adams, H. Wallach and Z. Ghahramani*
10:35 - 13:00	Coffee Break and Poster Session II
13:00 - 17:00	Lunch
	Approximate Inference
17:00 - 17:25	Solving the uncapacitated facility location problem using message passing problems *N. Lazic, B. Frey and P. Arabi*
17:25 - 17:50	Dense message passing for sparse principal component analysis *K. Sharp and M. Rattray*
17:50 - 18:15	Focused belief propagation for query-specific inference *A. Chechetka and C. Guestrin*
18:15 - 18:45	Tea Break

Online Learning, Control and Information Theory

18:45 - 19:10 Exploiting feature covariance in high-dimensional online learning
J. Ma, A. Kulesza, M. Dredze, K. Crammer, L. Saul and F. Pereira

19:10 - 19:35 REGO: Rank-based estimation of Renyi information using Euclidean graph optimization
B. Poczos, C. Szepesvari and S. Kirshner

19:35 - 20:00 Coherent inference on optimal play in game trees
P. Hennig, D. Stern and T. Graepel

20:00 - 22:00 Dinner

Saturday 15 May

07:45 - 08:45 Breakfast

Invited Talk

08:45 - 09:45 Nonparametric Learning of Functions and Graphs in High Dimensions
John Lafferty

Kernel Methods

09:45 - 10:10 Nonlinear functional regression: a functional RKHS approach
H. Kadri

10:10 - 10:35 On the relation between universality, characteristic kernels and RKHS embedding of measures
B. Sriperumbudur, K. Fukumizu and G. Lanckreit

10:35 - 13:00 Coffee Break and Poster Session III

13:00 - 17:00 Lunch

Graphical Models and Causal Inference

17:00 - 17:25 On combining graph-based variance reduction schemes
V. Gogate and R. Dechter

17:25 - 17:50 Convex structure learning in log-linear models beyond pairwise potentials
M. Schmidt and K. Murphy

17:50 - 18:15 Modeling annotator expertise: learning when everybody knows a bit of something
R. Rosales, Y. Yan, G. Fung and J. Dy

18:15 - 18:45	Tea Break
	Low-rank Methods and Information Retrieval
18:45 - 19:10	Fluid dynamics models for low rank discriminant analysis *Y.-K. Noh, B.-T. Zhang and D. Lee*
19:10 - 19:35	Reduced-rank hidden Markov models *S. Siddiqi, B. Boots and G. Gordon*
19:35 - 20:00	Half transductive ranking *B. Bai, J. Weston, D. Grangier, R. Collobert, C. Cortes and M. Mohri*
20:00 - 22:00	Dinner

Abstracts of papers

Oral sessions: Thursday, May 13th

Invited Talk

08:45 - 09:45 Forensic statistics: where are we and where are we going?
Richard Gill
I will discuss the present situation of Forensic statistics. Rapid developments in forensic science are putting statistics and probability more and more into the court-room lime-light, often with apalling results. Why is this and where should we go? Standard Bayesian and standard frequentist statistics are based on the wrong paradigms. Forensic statisticians have to learn from the learning community. But in forensic statistics, N=1. How can we learn?

Network Models

09:45 - 10:10 Boosted optimization for network classification
T. Hancock and H. Mamitsuka

In this paper we propose a new classification algorithm designed for application on complex networks motivated by algorithmic similarities between boosting learning and message passing. We consider a network classifier as a logistic regression where the variables define the nodes and the interaction effects define the edges. From this definition we represent the problem as a factor graph of local exponential loss functions. Using the factor graph representation it is possible to interpret the network classifier as an ensemble of individual node classifiers. We then combine ideas from boosted learning with network optimization algorithms to define two novel algorithms, Boosted Expectation Propagation (BEP) and Boosted Message Passing (BMP). These algorithms optimize the global network classifier performance by locally weighting each node classifier by the error of the surrounding network structure. We compare the performance of BEP and BMP to logistic regression as well state of the art penalized logistic regression models on simulated grid structured networks. The results show that using local boosting to optimize the performance of a network classifier increases classification performance and is especially powerful in cases when the whole network structure must be considered for accurate classification.

10:10 - 10:35 Detecting weak but hierarchically-structured patterns in networks
A. Singh and R. Nowak

The ability to detect weak distributed activation patterns in networks is critical to several applications, such as identifying the onset of anomalous activity or incipient congestion in the Internet, or faint traces of a biochemical spread by a sensor network. This is a challenging problem since weak distributed patterns can be invisible in per node statistics as well as a global network-wide aggregate. Most prior work considers situations in which the activation/non-activation of each node is statistically independent, but this is unrealistic in many problems. In this paper, we consider structured patterns arising from statistical dependencies in the activation process. Our contributions are three-fold. First, we propose a sparsifying transform that succinctly represents structured activation patterns that conform to a hierarchical dependency graph. Second, we establish that the proposed transform facilitates detection of very weak activation patterns that cannot be detected with existing methods. Third, we show that the structure of the hierarchical dependency graph governing the activation process, and hence the network transform, can be learnt from very few (logarithmic in network size) independent snapshots of network activity.

Statistical Learning Theory

17:00 - 17:25 Risk bounds for transduction and semi-supervised learning relative to data structure
G. Lever

We relate function class complexity to structure in the function domain. This facilitates risk analysis relative to cluster structure in the input space which is particularly effective in semi-supervised learning. In particular we quantify the complexity of function classes defined over a graph in terms of the graph structure.

17:25 - 17:50 Multiclass-multilabel classification with more labels than examples
O. Dekel and O. Shamir

We discuss multiclass-multilabel classification problems in which the set of possible labels is extremely large. Most existing multiclass-multilabel learning algorithms expect to observe a reasonably large sample from each class, and fail if they receive only a handful of examples with a given label. We propose and analyze the following two-stage approach: first use an arbitrary (perhaps heuristic) classification algorithm to construct an initial classifier, then apply a simple but principled method to augment this classifier by removing harmful labels from its output. A careful theoretical analysis allows us to justify our approach under some reasonable conditions (such as label sparsity and power-law distribution of label frequencies), even when the training set does not provide a statistically accurate representation of most classes. Surprisingly, our theoretical analysis continues to hold even when the number of classes exceeds the sample size.

We demonstrate the merits of our approach on the ambitious task of categorizing the entire web using the 1.5 million categories defined on Wikipedia.

17:50 - 18:15 Empirical Bernstein boosting
P. Shivaswamy and T. Jebara

Concentration inequalities that incorporate variance information (such as Bernstein's or Bennett's inequality) are often significantly tighter than counterparts (such as Hoeffding's inequality) that disregard variance. Nevertheless, many state of the art machine learning algorithms for classification problems like AdaBoost and support vector machines (SVMs) extensively use Hoeffding's inequalities to justify empirical risk minimization and its variants. This article proposes a novel boosting algorithm based on a recently introduced principle–sample variance penalization–which is motivated from an empirical version of Bernstein's inequality. This framework leads to an efficient algorithm that is as easy to implement as AdaBoost while producing a strict generalization. Experiments on a large number of datasets show significant performance gains over AdaBoost. This paper shows that sample variance penalization could be a viable alternative to empirical risk minimization.

Bayesian nonparametrics and causal inference

18:45 - 19:10 Sufficient covariates and linear propensity analysis
H. Guo and P. Dawid

Working within the decision-theoretic framework for causal inference, we study the properties of "sufficient covariates", which support causal inference from observational data, and possibilities for their reduction. In particular we illustrate the role of a propensity variable by means of a simple model, and explain why such a reduction typically does not increase (and may reduce) estimation efficiency.

19:10 - 19:35 Dirichlet process mixtures of generalised linear models
L. Hannah, D. Blei and W. Powell

We propose Dirichlet Process mixtures of Generalized Linear Models (DP-GLMs), a new method of nonparametric regression that accommodates continuous and categorical inputs, models a response variable locally by a generalized linear model. We give conditions for the existence and asymptotic unbiasedness of the DP-GLM regression mean function estimate; we then give a practical example for when those conditions hold. We evaluate DP-GLM on several data sets, comparing it to modern methods of nonparametric regression including regression trees and Gaussian processes.

19:35 - 20:00 Bayesian Gaussian process latent variable model
M. Titsias and N. Lawrence

We introduce a variational inference framework for training the Gaussian process latent variable model and thus performing Bayesian nonlinear dimensionality reduction. This method allows us to variationally integrate out the input variables of the Gaussian process and compute a lower bound on the exact marginal likelihood of the nonlinear latent variable model. The maximization of the variational lower bound provides a Bayesian training procedure that is robust to overfitting and can automatically select the dimensionality of the nonlinear latent space. We demonstrate our method on real world datasets. The focus in this paper is on dimensionality reduction problems, but the methodology is more general. For example, our algorithm is immediately applicable for training Gaussian process models in the presence of missing or uncertain inputs.

Poster session I: Thursday, May 13th (10:35 - 13:00)

Contributed Posters

TP1 Online Anomaly Detection under Adversarial Impact
M. Kloft and P. Laskov

Security analysis of learning algorithms is gaining increasing importance, especially since they have become target of deliberate obstruction in certain applications. Some security-hardened algorithms have been previously proposed for supervised learning; however, very little is known about the behavior of anomaly detection methods in such scenarios. In this contribution, we analyze the performance of a particular method—online centroid anomaly detection—in the presence of adversarial noise. Our analysis addresses three key security-related issues: derivation of an optimal attack, analysis of its efficiency and constraints. Experimental evaluation carried out on real HTTP and exploit traces confirms the tightness of our theoretical bounds.

TP2 A Weighted Multi-Sequence Markov Model For Brain Lesion Segmentation
F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot and M. Dojat

We propose a technique for fusing the output of multiple Magnetic Resonance (MR) sequences to robustly and accurately segment brain lesions. It is based on an augmented multi-sequence Hidden Markov model that includes additional weight variables to account for the relative importance and control the impact of each sequence. The augmented framework has the advantage of allowing 1) the incorporation of expert knowledge on the a priori relevant information content of each sequence and 2) a weighting scheme which is modified adaptively according to the data and the segmentation task under consideration. The model, applied to the detection of multiple sclerosis and stroke lesions shows promising results.

TP3 Online Passive-Aggressive Algorithms on a Budget
Z. Wang and S. Vucetic

In this paper a kernel-based online learning algorithm, which has both constant space and update time, is proposed. The approach is based on the popular online Passive-Aggressive (PA) algorithm. When used in conjunction with kernel function, the number of support vectors in PA grows without bounds when learning from noisy data streams. This implies unlimited memory and ever increasing model update and prediction time. To address this issue, the proposed budgeted PA algorithm maintains only a fixed number of support vectors. By introducing an additional constraint to the original PA optimization problem, a closed-form solution was derived for the support vector removal and model update. Using the hinge loss we developed several budgeted PA algorithms that can trade between accuracy and update cost. We also developed the ramp loss versions of both original and budgeted PA and showed that the resulting algorithms can be interpreted as the combination of active learning and hinge loss PA. All proposed algorithms were comprehensively tested on 7 benchmark data sets. The experiments showed that they are superior to the existing budgeted online algorithms. Even with modest budgets, the budgeted PA achieved very competitive accuracies to the non-budgeted PA and kernel perceptron algorithms.

TP4 Guarantees for Approximate Incremental SVMs
N. Usunier, A. Bordes and L. Bottou

Assume a teacher provides examples one by one. An approximate incremental SVM computes a sequence of classifiers that are close to the true SVM solutions computed on the successive incremental training sets. We show that simple algorithms can satisfy an averaged accuracy criterion with a computational cost that scales as well as the best SVM algorithms with the number of examples. Finally, we exhibit some experiments highlighting the benefits of joining fast incremental optimization and curriculum and active learning (Schon and Cohn, 2000; Bordes et al., 2005; Bengio et al., 2009).

TP5 Near-Optimal Evasion of Convex-Inducing Classifiers
B. Nelson, B. Rubinstein, L. Huang, A. Joseph, S. Lau, S. Lee, S. Rao, A. Tran and D. Tygar

Classifiers are often used to detect miscreant activities. We study how an adversary can efficiently query a classifier to elicit information that allows the adversary to evade detection at near-minimal cost. We generalize results of Lowd and Meek (2005) to convex-inducing classifiers. We present algorithms that construct undetected instances of near-minimal cost using only polynomially many queries in the dimension of the space and without reverse engineering the decision boundary.

TP6 A Regularization Approach to Nonlinear Variable Selection
L. Rosasco, M. Santoro, S. Mosci, A. Verri and S. Villa

In this paper we consider a regularization approach to variable selection when the regression function depends nonlinearly on a few input variables. The proposed method is based on a regularized least square estimator penalizing large values of the partial derivatives. An efficient iterative procedure is proposed to solve the underlying variational problem, and its convergence is proved. The empirical properties of the obtained estimator are tested both for prediction and variable selection. The algorithm compares favorably to more standard ridge regression and L1 regularization schemes.

TP7 Exploiting Covariate Similarity in Sparse Regression via the Pairwise Elastic Net
A. Lorbert, D. Eis, V. Kostina, D. Blei and P. Ramadge

A new approach to regression regularization called the Pairwise Elastic Net is proposed. Like the Elastic Net, it simultaneously performs automatic variable selection and continuous shrinkage. In addition, the Pairwise Elastic Net encourages the grouping of strongly correlated predictors based on a pairwise similarity measure. We give examples of how the Pairwise Elastic Net can be used to achieve the objectives of Ridge regression, the Lasso, the Elastic Net, and Group Lasso. Finally, we present a coordinate descent algorithm to solve the Pairwise Elastic Net.

TP8 The Group Dantzig Selector
H. Liu, J. Zhang, X. Jiang and J. Liu

We introduce a new method – the group Dantzig selector – for high dimensional sparse regression with group structure, which has a convincing theory about why utilizing the group structure can be beneficial. Under a group restricted isometry condition, we obtain a significantly improved nonasymptotic L2-norm bound over the basis pursuit or the Dantzig selector which ignores the group structure. To gain more insight, we also introduce a surprisingly simple and intuitive "sparsity oracle condition" to obtain a block L1-norm bound, which is easily accessible to a broad audience in machine learning community. Encouraging numerical results are also provided to support our theory.

TP9 Ultra-high Dimensional Multiple Output Learning With Simultaneous Orthogonal Matching Pursuit: Screening Approach
M. Kolar and E. Xing

We propose a novel application of the Simultaneous Orthogonal Matching Pursuit (S-OMP) procedure to perform variable selection in ultra-high dimensional multiple output regression problems, which is the first attempt to utilize multiple outputs to perform fast removal of the irrelevant variables. As our main theoretical contribution, we show that the S-OMP can be used to reduce an ultra-high number of variables to below the sample size, without losing relevant variables. We also provide formal evidence that the modified Bayesian information criterion (BIC) can be used to efficiently select the number of iterations in the S-OMP. Once the number of variables has been reduced to a manageable size, we show that a more computationally demanding procedure can be used to identify the relevant variables for each of the regression outputs.

We further provide evidence on the benefit of variable selection using the regression outputs jointly, as opposed to performing variable selection for each output separately. The finite sample performance of the S-OMP has been demonstrated on extensive simulation studies.

TP10 The Feature Selection Path in Kernel Methods
F. Li and C. Sminchisescu

The problem of automatic feature selection/weighting in kernel methods is examined. We work on a formulation that optimizes both the weights of features and the parameters of the kernel model simultaneously, using L_1 regularization for feature selection. Under quite general choices of kernels, we prove that there exists a unique regularization path for this problem, that runs from 0 to a stationary point of the non-regularized problem. We propose an ODE-based homotopy method to follow this trajectory. By following the path, our algorithm is able to automatically discard irrelevant features and to automatically go back and forth to avoid local optima. Experiments on synthetic and real datasets show that the method achieves low prediction error and is efficient in separating relevant from irrelevant features.

TP11 Exclusive Lasso for Multi-task Feature Selection
Y. Zhou, R. Jin and S. Chu-Hong Hoi

We propose a novel group regularization which we call exclusive lasso. Unlike the group lasso regularizer that assumes co-varying variables in groups, the proposed exclusive lasso regularizer models the scenario when variables in the same group compete with each other. Analysis is presented to illustrate the properties of the proposed regularizer. We present a framework of kernel-based multi-task feature selection algorithm based on the proposed exclusive lasso regularizer. An efficient algorithm is derived to solve the related optimization problem. Experiments with document categorization show that our approach outperforms state-of-the-art algorithms for multi-task feature selection.

TP12 Semi-Supervised Learning via Generalized Maximum Entropy
A. Erkan and Y. Altun

Various supervised inference methods can be analyzed as convex duals of the generalized maximum entropy (MaxEnt) framework. Generalized MaxEnt aims to find a distribution that maximizes an entropy function while respecting prior information represented as potential functions in miscellaneous forms of constraints and/or penalties. We extend this framework to semi-supervised learning by incorporating unlabeled data via modifications to these potential functions reflecting structural assumptions on the data geometry. The proposed approach leads to a family of discriminative semi-supervised algorithms, that are convex, scalable, inherently multi-class, easy to implement, and that can be kernelized naturally. Experimental evaluation of special cases shows the competitiveness of our methodology.

TP13 Semi-Supervised Learning with Max-Margin Graph Cuts
B. Kveton, M. Valko, A. Rahimi and L. Huang

This paper proposes a novel algorithm for semi-supervised learning. This algorithm learns graph cuts that maximize the margin with respect to the labels induced by the harmonic function solution. We motivate the approach, compare it to existing work, and prove a bound on its generalization error. The quality of our solutions is evaluated on a synthetic problem and three UCI ML repository datasets. In most cases, we outperform manifold regularization of support vector machines, which is a state-of-the-art approach to semi-supervised max-margin learning.

TP14 Bayesian variable order Markov models
C. Dimitrakakis

We present a simple, effective generalisation of variable order Markov models to full online Bayesian estimation. The mechanism used is close to that employed in context tree weighting. The main contribution is the addition of a prior, conditioned on context, on the Markov order. The resulting construction uses a simple recursion and can be updated efficiently. This allows the model to make predictions using more complex contexts, as more data is acquired, if necessary. In addition, our model can be alternatively seen as a mixture of tree experts. Experimental results show that the predictive model exhibits consistently good performance in a variety of domains.

TP15 Bayesian Generalized Kernel Models
Z. Zhang, G. Dai, D. Wang and M. Jordan

We propose a fully Bayesian approach for generalized kernel models (GKMs), which are extensions of generalized linear models in the feature space induced by a reproducing kernel. We place a mixture of a point-mass distribution and Silverman's g-prior on the regression vector of GKMs. This mixture prior allows a fraction of the regression vector to be zero. Thus, it serves for sparse modeling and Bayesian computation. For inference, we exploit data augmentation methodology to develop a Markov chain Monte Carlo (MCMC) algorithm in which the reversible jump method is used for model selection and a Bayesian model averaging method is used for posterior prediction.

TP16 Mass Fatality Incident Identification based on nuclear DNA evidence
F. Corradi

This paper focuses on the use of nuclear DNA Short Tandem Repeat traits for the identification of the victims of a Mass Fatality Incident. The goal of the analysis is the assessment of the identification probabilities concerning the recovered victims. Identification hypotheses are evaluated conditionally to the DNA evidence observed both on the recovered victims and on the relatives of the missing persons disappeared in the tragical event. After specifying a set of conditional independence assertions suitable for the problem, an inference strategy is provided, treating some points to achieve computational efficiency. Finally, the proposal is tested through the simulation of a Mass Fatality Incident and the results are examined in details.

TP17 Approximate parameter inference in a stochastic reaction-diffusion model
A. Ruttor and M. Opper

We present an approximate inference approach to parameter estimation in a spatio-temporal stochastic process of the reaction-diffusion type. The continuous space limit of an inference method for Markov jump processes leads to an approximation which is related to a spatial Gaussian process. An efficient solution in feature space using a Fourier basis is applied to inference on simulational data.

TP18 Parametric Herding
Y. Chen and M. Welling

A parametric version of herding is formulated. The nonlinear mapping between consecutive time slices is learned by a form of self-supervised training. The resulting dynamical system generates pseudo-samples that resemble the original data. We show how this parametric herding can be successfully used to compress a dataset consisting of binary digits. It is also verified that high compression rates translate into good prediction performance on unseen test data.

TP19 Noise-contrastive estimation: A new estimation principle for unnormalized statistical models
M. Gutmann and A. Hyvärinen

We present a new estimation principle for parameterized statistical models. The idea is to perform nonlinear logistic regression to discriminate between the observed data and some artificially generated noise, using the model log-density function in the regression nonlinearity. We show that this leads to a consistent (convergent) estimator of the parameters, and analyze the asymptotic variance. In particular, the method is shown to directly work for unnormalized models, i.e. models where the density function does not integrate to one. The normalization constant can be estimated just like any other parameter. For a tractable ICA model, we compare the method with other estimation methods that can be used to learn unnormalized models, including score matching, contrastive divergence, and maximum-likelihood where the normalization constant is estimated with importance sampling. Simulations show that noise-contrastive estimation offers the best trade-off between computational and statistical efficiency. The method is then applied to the modeling of natural images: We show that the method can successfully estimate a large-scale two-layer model and a Markov random field.

TP20 Understanding the difficulty of training deep feedforward neural networks
X. Glorot and Y. Bengio

Whereas before 2006 it appears that deep multi-layer neural networks were not successfully trained, since then several algorithms have been shown to successfully train them, with experimental results showing the superiority of deeper vs less deep architectures. All these experimental results were obtained with new initialization or training mechanisms. Our objective here is to understand better why standard gradient descent from random initialization is doing so poorly with deep neural networks, to better understand these recent relative successes and help design better algorithms in the future. We first observe the influence of the non-linear activations functions. We find that the logistic sigmoid activation is unsuited for deep networks with random initialization because of its mean value, which can drive especially the top hidden layer into saturation. Surprisingly, we find that saturated units can move out of saturation by themselves, albeit slowly, and explaining the plateaus sometimes seen when training neural networks. We find that a new non-linearity that saturates less can often be beneficial. Finally, we study how activations and gradients vary across layers and during training, with the idea that training may be more difficult when the singular values of the Jacobian associated with each layer are far from 1. Based on these considerations, we propose a new initialization scheme that brings substantially faster convergence.

TP21 Exploiting Within-Clique Factorizations in Junction-Tree Algorithms
J. McAuley and T. Caetano

We show that the expected computational complexity of the Junction-Tree Algorithm for maximum a posteriori inference in graphical models can be improved. Our results apply whenever the potentials over maximal cliques of the triangulated graph are factored over subcliques. This is common in many real applications, as we illustrate with several examples. The new algorithms are easily implemented, and experiments show substantial speed-ups over the classical Junction-Tree Algorithm. This enlarges the class of models for which exact inference is efficient.

TP22 Maximum-likelihood learning of cumulative distribution functions on graphs
J. Huang and N. Jojic

For many applications, a probability model can be easily expressed as a cumulative distribution functions (CDF) as compared to the use of probability density or mass functions (PDF/PMFs). Cumulative distribution networks (CDNs) have recently been proposed as a class of graphical models for CDFs. One advantage of CDF models is the simplicity of representing multivariate heavy-tailed distributions. Examples of fields that can benefit from the use of graphical models for CDFs include climatology and epidemiology, where data may follow extreme value statistics and exhibit spatial correlations so that dependencies between model variables must be accounted for. The problem of learning from data in such settings may nevertheless consist of optimizing the log-likelihood function with respect to model parameters where we are required to optimize a log-PDF/PMF and not a log-CDF. We present a message-passing algorithm called the gradient-derivative-product (GDP) algorithm that allows us to learn the model in terms of the log-likelihood function whereby messages correspond to local gradients of the likelihood with respect to model parameters.

We will demonstrate the GDP algorithm on real-world rainfall and H1N1 mortality data and we will show that CDNs provide a natural choice of parameterizations for the heavy-tailed multivariate distributions that arise in these problems.

TP23 Improving posterior marginal approximations in latent Gaussian models
B. Cseke and T. Heskes

We consider the problem of correcting the posterior marginal approximations computed by expectation propagation and Laplace approximation in latent Gaussian models and propose correction methods that are similar in spirit to the Laplace approximation of Tierney and Kadane (1986). We show that in the case of sparse Gaussian models, the computational complexity of expectation propagation can be made comparable to that of the Laplace approximation by using a parallel updating scheme. In some cases, expectation propagation gives excellent estimates, where the Laplace approximation fails. Inspired by bounds on the marginal corrections, we arrive at factorized approximations, which can be applied on top of both expectation propagation and Laplace. These give nearly indistinguishable results from the non-factorized approximations in a fraction of the time.

TP24 Nonparametric Tree Graphical Models
L. Song, A. Gretton and C. Guestrin

We introduce a nonparametric representation for graphical model on trees which expresses marginals as Hilbert space embeddings and conditionals as embedding operators. This formulation allows us to define a graphical model solely on the basis of the feature space representation of its variables. Thus, this nonparametric model can be applied to general domains where kernels are defined, handling challenging cases such as discrete variables whose domains are huge, or very complex, non-Gaussian continuous distributions. We also derive *kernel belief propagation*, a Hilbert-space algorithm for performing inference in our model. We show that our method outperforms state-of-the-art techniques in a cross-lingual document retrieval task and a camera rotation estimation problem.

TP25 Structured Sparse Principal Component Analysis
R. Jenatton, g. Obozinski and F. Bach

We present an extension of sparse PCA, or sparse dictionary learning, where the sparsity patterns of all dictionary elements are structured and constrained to belong to a prespecified set of shapes. This structured sparse PCA is based on a structured regularization recently introduced by Jenatton et al.(2009). While classical sparse priors only deal with cardinality, the regularization we use encodes higher-order information about the data. We propose an efficient and simple optimization procedure to solve this problem. Experiments with two practical tasks, the denoising of sparse structured signals and face recognition, demonstrate the benefits of the proposed structured approach over unstructured approaches.

TP26 Factorized Orthogonal Latent Spaces
M. Salzmann, C. Henrik Ek, R. Urtasun and T. Darrell

Existing approaches to multi-view learning are particularly effective when the views are either independent (i.e, multi-kernel approaches) or fully dependent (i.e., shared latent spaces). However, in real scenarios, these assumptions are almost never truly satisfied. Recently, two methods have attempted to tackle this problem by factorizing the information and learn separate latent spaces for modeling the shared (i.e., correlated) and private (i.e., independent) parts of the data. However, these approaches are very sensitive to parameters setting or initialization. In this paper we propose a robust approach to factorizing the latent space into shared and private spaces by introducing orthogonality constraints, which penalize redundant latent representations. Furthermore, unlike previous approaches, we simultaneously learn the structure and dimensionality of the latent spaces by relying on a regularizer that encourages the latent space of each data stream to be low dimensional. To demonstrate the benefits of our approach, we apply it to two existing shared latent space models that assume full dependence of the views, the sGPLVM and the sKIE, and show that our constraints improve the performance of these models on the task of pose estimation from monocular images.

TP27 Sufficient Dimension Reduction via Squared-loss Mutual Information Estimation
T. Suzuki and M. Sugiyama

The goal of sufficient dimension reduction in supervised learning is to find the lowdimensional subspace of input features that is sufficient for predicting output values. In this paper, we propose a novel sufficient dimension reduction method using a squaredloss variant of mutual information as a dependency measure. We utilize an analytic approximator of squared-loss mutual information based on density ratio estimation, which is shown to possess suitable convergence properties. We then develop a natural gradient algorithm for sufficient subspace search. Numerical experiments show that the proposed method compares favorably with existing dimension reduction approaches.

TP28 Identifying Cause and Effect on Discrete Data using Additive Noise Models
J. Peters, D. Janzing and B. Schoelkopf

Inferring the causal structure of a set of random variables from a finite sample of the joint distribution is an important problem in science. Recently, methods using additive noise models have been suggested to approach the case of continuous variables. In many situations, however, the variables of interest are discrete or even have only finitely many states. In this work we extend the notion of additive noise models to these cases. Whenever the joint distribution $P^{(}X, Y)$ admits such a model in one direction, e.g. Y=f(X)+N, N independent of X, it does not admit the reversed model X=g(Y)+N', N' independent of Y as long as the model is chosen in a generic way. Based on these deliberations we propose an efficient new algorithm that is able to distinguish between cause and effect for a finite sample of discrete variables. We show that this algorithm works both on synthetic and real data sets.

TP29 Using Descendants as Instrumental Variables for the Identification of Direct Causal Effects in Linear SEMs

H. Chan and M. Kuroki

In this paper, we present an extended set of graphical criteria for the identification of direct causal effects in linear Structural Equation Models (SEMs). Previous methods of graphical identification of direct causal effects in linear SEMs include methods such as the single-door criterion, the instrumental variable and the IV-pair, and the accessory set. However, there remain graphical models where a direct causal effect can be identified and these graphical criteria all fail. As a result, we introduce a new set of graphical criteria which uses descendants of either the cause variable or the effect variable as “path-specific instrumental variables” for the identification of the direct causal effect as long as certain conditions are satisfied. These conditions are based on edge removal and the existing graphical criteria of instrumental variables, and the identifiability of certain other total effects, and thus can be easily checked.

TP30 Learning Causal Structure from Overlapping Variable Sets
S. Triantafillou, I. Tsamardinos and I. Tollis

We present an algorithm name cSAT+ for learning the causal structure in a domain from datasets measuring different variables sets. The algorithm outputs a graph with edges corresponding to all possible pairwise causal relations between two variables, named Pairwise Causal Graph (PCG). Examples of interesting inferences include the induction of the absence or presence of some causal relation between two variables never measured together. cSAT+ converts the problem to a series of SAT problems, obtaining leverage from the efficiency of state-of-the-art solvers. In our empirical evaluation, it is shown to outperform ION, the first algorithm solving a similar but more general problem, by two orders of magnitude.

TP31 Combining Experiments to Discover Linear Cyclic Models with Latent Variables
F. Eberhardt, P. Hoyer and R. Scheines

We present an algorithm to infer causal relations between a set of measured variables on the basis of experiments on these variables. The algorithm assumes that the causal relations are linear, but is otherwise completely general: It provides consistent estimates when the true causal structure contains feedback loops and latent variables, while the experiments can involve surgical or ’soft’ interventions on one or multiple variables at a time. The algorithm is ’online’ in the sense that it combines the results from any set of available experiments, can incorporate background knowledge and resolves conflicts that arise from combining results from different experiments. In addition we provide a necessary and sufficient condition that (i) determines when the algorithm can uniquely return the true graph, and (ii) can be used to select the next best experiment until this condition is satisfied. We demonstrate the method by applying it to simulated data and the flow cytometry data of Sachs et al (2005).

TP32 Inference and Learning in Networks of Queues
C. Sutton and M. Jordan

Probabilistic models of the performance of computer systems are useful both for predicting system performance in new conditions, and for diagnosing past performance problems. The most popular performance models are networks of queues. However, no current methods exist for parameter estimation or inference in networks of queues with missing data. In this paper, we present a novel viewpoint that combines queueing networks and graphical models, allowing Markov chain Monte Carlo to be applied. We demonstrate the effectiveness of our sampler on real-world data from a benchmark Web application.

TP33 Deterministic Bayesian inference for the p* model
H. Austad and N. Friel

The p* model is widely used in social network analysis. The likelihood of a network under this model is impossible to calculate for all but trivially small networks. Various approximation have been presented in the literature, and the pseudolikelihood approximation is the most popular. The aim of this paper is to introduce two likelihood approximations which have the pseudolikelihood estimator as a special case. We show, for the examples that we have considered, that both approximations result in improved estimation of model parameters with respect to the standard methodological approaches. We provide a deterministic approach and also illustrate how Bayesian model choice can be carried out in this setting.

Posters from Breaking-News Abstracts

TA1 Have I seen you before? Principles of Bayesian predictive classification revisited
J. Corander, Y. Cui, T. Koski and J. Siren

TA2 Decisive symmetric games: study of their decisiveness
F. Carreras, J. Freixas and M.A. Puente

TA3 A new variational Bayesian algorithm with Gaussian mixture component spitting for mining spatial data
B. Wu, C.A. McGrory and A.N. Pettitt

TA4 Variables selection in unsupervised classification by mixture using genotype data
W. Toussile

TA5 A convex regularization formulation for learning task relationships in multi-task learning
Y. Zhang and D.-Y. Yeung

TA6 Recursive modeling using xstatR
E.J. Harner and J. Tan

TA7 Adaptation and complexity regularisation in data streams
N.G. Pavlidis, D.K. Tasoulis, N.M. Adams and D.J. Hand

TA8 A general mixed-membership model with application to children's learning
A. Galyardt

TA9 Adaptive estimation for multivariate Gaussian data-streams: an approximately Bayesian approach
P. Rubin-Delanchy

TA10 Results of the active learning challenge
I. Guyon, G. Cawley, G. Dror and V, Lemaire

TA11 Learning why things change: the difference-based causality learner
M. Voortman, D. Dash and M.J. Druzdzel

TA12 A model-based method for transcription factor target identification from short gene expression time series data
A. Honkela, N.D. Lawrence and M. Rattray

Oral sessions: Friday, May 14th

Invited Talk

08:45 - 09:45 Approximate Bayesian Computation: What, Why and How?
Simon Tavaré

Approximate Bayesian Computation (ABC) arose in response to the difficulty of simulating observations from posterior distributions determined by intractable likelihoods. The method exploits the fact that while likelihoods may be impossible to compute in complex probability models, it is often easy to simulate observations from them. ABC in its simplest form proceeds as follows: (i) simulate a parameter from the prior; (ii) simulate observations from the model with this parameter; (iii) accept the parameter if the simulated observations are close enough to the observed data. The magic, and the source of potential disasters, is in step (iii). This talk will outline what we know (and don't!) about ABC and illustrate the methods with applications to the fossil record and stem cell biology.

Deep Learning

09:45 - 10:10 Factored 3-way restricted Boltzmann machines for modeling natural images
M. Ranzato, A. Krizhevsky and G. Hinton

Deep belief nets have been successful in modeling handwritten characters, but it has proved more difficult to apply them to real images. The problem lies in the restricted Boltzmann machine (RBM) which is used as a module for learning deep belief nets one layer at a time. The Gaussian-Binary RBMs that have been used to model real-valued data are not a good way to model the covariance structure of natural images. We propose a factored 3-way RBM that uses the states of its hidden units to represent abnormalities in the local covariance structure of an image. This provides a probabilistic framework for the widely used simple/complex cell architecture. Our model learns binary features that work very well for object recognition on the "tiny images" data set. Even better features are obtained by then using standard binary RBM's to learn a deeper model.

10:10 - 10:35 Learning the structure of deep sparse graphical models
R. Adams, H. Wallach and Z. Ghahramani

Deep belief networks are a powerful way to model complex probability distributions. However, it is difficult to learn the structure of a belief network, particularly one with hidden units. The Indian buffet process has been used as a nonparametric Bayesian prior on the structure of a directed belief network with a single infinitely wide hidden layer. Here, we introduce the cascading Indian buffet process (CIBP), which provides a prior on the structure of a layered, directed belief network that is unbounded in both depth and width, yet allows tractable inference. We use the CIBP prior with the nonlinear Gaussian belief network framework to allow each unit to vary its behavior between discrete and continuous representations. We use Markov chain Monte Carlo for inference in this model and explore the structures learned on image data.

Approximate Inference

17:00 - 17:25 Solving the uncapacitated facility location problem using message passing problems
N. Lazic, B. Frey and P. Arabi

The Uncapacitated Facility Location Problem (UFLP) is one of the most widely studied discrete location problems, whose applications arise in a variety of settings. We tackle the UFLP using probabilistic inference in a graphical model - an approach that has received little attention in the past. We show that the fixed points of max-product linear programming (MPLP), a convexified version of the max-product algorithm, can be used to construct a solution with a 3-approximation guarantee for metric UFLP instances. In addition, we characterize some scenarios under which the MPLP solution is guaranteed to be globally optimal. We evaluate the performance of both max-sum and MPLP empirically on metric and non-metric problems, demonstrating the advantages of the 3-approximation construction and algorithm applicability to non-metric instances.

17:25 - 17:50 Dense message passing for sparse principal component analysis
K. Sharp and M. Rattray

We describe a novel inference algorithm for sparse Bayesian PCA with a zero-norm prior on the model parameters. Bayesian inference is very challenging in probabilistic models of this type. MCMC procedures are too slow to be practical in a very high-dimensional setting and standard mean-field variational Bayes algorithms are ineffective. We adopt a dense message passing algorithm similar to algorithms developed in the statistical physics community and previously applied to inference problems in coding and sparse classification. The algorithm achieves near-optimal performance on synthetic data for which a statistical mechanics theory of optimal learning can be derived. We also study two gene expression datasets used in previous studies of sparse PCA. We find our method performs better than one published algorithm and comparably to a second.

17:50 - 18:15 Focused belief propagation for query-specific inference
A. Chechetka and C. Guestrin

With the increasing popularity of large-scale probabilistic graphical models, even "lightweight" approximate inference methods are becoming infeasible. Fortunately, often large parts of the model are of no immediate interest to the end user. Given the variable that the user actually cares about, we show how to quantify edge importance in graphical models and to significantly speed up inference by focusing computation on important parts of the model. Our algorithm empirically demonstrates convergence speedup by multiple times over state of the art

Online Learning, Control and Information Theory

18:45 - 19:10 Exploiting feature covariance in high-dimensional online learning
J. Ma, A. Kulesza, M. Dredze, K. Crammer, L. Saul and F. Pereira

Some online algorithms for linear classification model the uncertainty in their weights over the course of learning. Modeling the full covariance structure of the weights can provide a significant advantage for classification. However, for high-dimensional, large-scale data, even though there may be many second-order feature interactions, it is computationally infeasible to maintain this covariance structure. To extend second-order methods to high-dimensional data, we develop low-rank approximations of the covariance structure. We evaluate our approach on both synthetic and real-world data sets using the confidence-weighted online learning framework. We show improvements over diagonal covariance matrices for both low and high-dimensional data.

19:10 - 19:35 REGO: Rank-based estimation of Renyi information using Euclidean graph optimization
B. Poczos, C. Szepesvari and S. Kirshner

We propose a new method for a non-parametric estimation of Renyi and Shannon information for a multivariate distribution using a corresponding copula, a multivariate distribution over normalized ranks of the data. As the information of the distribution is the same as the negative entropy of its copula, our method estimates this information by solving a Euclidean graph optimization problem on the empirical estimate of the distribution's copula. Owing to the properties of the copula, we show that the resulting estimator of Renyi information is strongly consistent and robust. Further, we demonstrate its applicability in the image registration in addition to simulated experiments.

19:35 - 20:00 Coherent inference on optimal play in game trees
P. Hennig, D. Stern and T. Graepel

Round-based games are an instance of discrete planning problems. Some of the best contemporary game tree search algorithms use random roll-outs as data. Relying on a good policy, they learn on-policy values by propagating information upwards in the tree, but not between sibling nodes. Here, we present a generative model and a corresponding approximate message passing scheme for inference on the optimal, off-policy value of nodes in smooth AND/OR trees, given random roll-outs. The crucial insight is that the distribution of values in game trees is not completely arbitrary. We define a generative model of the on-policy values using a latent score for each state, representing the value under the random roll-out policy. Inference on the values under the optimal policy separates into an inductive, pre-data step and a deductive, post-data part. Both can be solved approximately with Expectation Propagation, allowing off-policy value inference for any node in the (exponentially big) tree in linear time.

Poster session II: Friday, May 14th (10:35 - 13:00)

Contributed Posters

FP1 On the Impact of Kernel Approximation on Learning Accuracy
C. Cortes, M. Mohri and A. Talwalkar

Kernel approximation is commonly used to scale kernel-based algorithms to applications containing as many as several million instances. This paper analyzes the effect of such approximations in the kernel matrix on the hypothesis generated by several widely used learning algorithms. We give stability bounds based on the norm of the kernel approximation for these algorithms, including SVMs, KRR, and graph Laplacian-based regularization algorithms. These bounds help determine the degree of approximation that can be tolerated in the estimation of the kernel matrix. Our analysis is general and applies to arbitrary approximations of the kernel matrix. However, we also give a specific analysis of the Nystrom low-rank approximation in this context and report the results of experiments evaluating the quality of the Nystrom low-rank kernel approximation when used with ridge regression.

FP2 Kernel Partial Least Squares is Universally Consistent
G. Blanchard and N. Krämer

We prove the statistical consistency of kernel Partial Least Squares Regression applied to a bounded regression learning problem on a reproducing kernel Hilbert space. Partial Least Squares stands out of well-known classical approaches as e.g. Ridge Regression or Principal Components Regression, as it is not defined as the solution of a global cost minimization procedure over a fixed model nor is it a linear estimator. Instead, approximate solutions are constructed by projections onto a nested set of data-dependent subspaces. To prove consistency, we exploit the known fact that Partial Least Squares is equivalent to the conjugate gradient algorithm in combination with early stopping. The choice of the stopping rule (number of iterations) is a crucial point. We study two empirical stopping rules. The first one monitors the estimation error in each iteration step of Partial Least Squares, and the second one estimates the empirical complexity in terms of a condition number. Both stopping rules lead to universally consistent estimators provided the kernel is universal.

FP3 Risk Bounds for Levy Processes in the PAC-Learning Framework
C. Zhang and D. Tao

Levy processes play an important role in the stochastic process theory. However, since samples are non-i.i.d., statistical learning results based on the i.i.d. scenarios cannot be utilized to study the risk bounds for Levy processes. In this paper, we present risk bounds for non-i.i.d. samples drawn from Levy processes in the PAC-learning framework. In particular, by using a concentration inequality for infinitely divisible distributions, we first prove that the function of risk error is Lipschitz continuous with a high probability, and then by using a specific concentration inequality for Levy processes, we obtain the risk bounds for non-i.i.d. samples drawn from Levy processes without Gaussian components. Based on the resulted risk bounds, we analyze the factors that affect the convergence of the risk bounds and then prove the convergence.

FP4 Negative Results for Active Learning with Convex Losses
S. Hanneke and L. Yang

We study the problem of active learning with convex loss functions. We prove that even under bounded noise constraints, the minimax rates for proper active learning are often no better than passive learning.

FP5 Impossibility Theorems for Domain Adaptation
S. Ben David, T. Lu, T. Luu and D. Pal

The domain adaptation problem in machine learning occurs when the test data generating distribution differs from the one that generates the training data. It is clear that the success of learning under such circumstances depends on similarities between the two data distributions. We study assumptions about the relationship between the two distributions that one needed for domain adaptation learning to succeed. We analyze the assumptions in an agnostic PAC-style learning model for a the setting in which the learner can access a labeled training data sample and an unlabeled sample generated by the test data distribution. We focus on three assumptions: (i) Similarity between the unlabeled distributions, (ii) Existence of a classifier in the hypothesis class with low error on both training and testing distributions, and (iii) The covariate shift assumption.I.e., the assumption that the conditioned label distribution (for each data point) is the same for both the training and test distributions. We show that without either assumption (i)or (ii), the combination of the remaining assumptions is not sufficient toguarantee successful learning. Our negative results hold with respect to any domain adaptation learning algorithm, as long as it does not have access to target labeled examples. In particular, we provide formal proofs that the popular covariate shift assumption is rather weak and does not relieve the necessity of the other assumptions. We also discuss the intuitively appealing paradigm of reweighing the labeled training sample according to the target unlabeled distribution. We show that, somewhat counter intuitively, that paradigm cannot be trusted in the following sense. There are DA tasks that are indistinguishable, as far as the input training data goes, but in which reweighing leads to significant improvement in one task, while causing dramatic deterioration of the learning success in the other.

FP6 Bayesian Online Learning for Multi-label and Multi-variate Performance Measures
X. Zhang, T. Graepel and R. Herbrich

Many real world applications employ multi-variate performance measures and each example can belong to multiple classes. The currently most popular approaches train an SVM for each class, followed by ad hoc thresholding. Probabilistic models using Bayesian decision theory are also commonly adopted. In this paper, we propose a Bayesian online multi-label classification framework (BOMC) which learns a probabilistic linear classifier. The likelihood is modeled by a graphical model similar to TrueSkill^{TM}, and inference is based on Gaussian density filtering with expectation propagation. Using samples from the posterior, we label the testing data by maximizing the expected F_1-score. Our experiments on Reuters1-v2 dataset show BOMC compares favorably to the state-of-the-art online learners in macro-averaged F_1-score and training time.

FP7 Infinite Predictor Subspace Models for Multitask Learning
P. Rai and H. Daume III

Given several related learning tasks, we propose a nonparametric Bayesian model that captures task relatedness by assuming that the task parameters (i.e., predictors) share a latent subspace. More specifically, the intrinsic dimensionality of the task subspace is not assumed to be known a priori. We use an infinite latent feature model to automatically infer this number (depending on and limited by only the number of tasks). Furthermore, our approach is applicable when the underlying task parameter subspace is inherently sparse, drawing parallels with l1 regularization and LASSO-style models. We also propose an augmented model which can make use of (labeled, and additionally unlabeled if available) inputs to assist learning this subspace, leading to further improvements in the performance. Experimental results demonstrate the efficacy of both the proposed approaches, especially when the number of examples per task is small. Finally, we discuss an extension of the proposed framework where a nonparametric mixture of linear subspaces can be used to learn a manifold over the task parameters, and also deal with the issue of negative transfer from unrelated tasks.

FP8 Neural conditional random fields
T. Do and T. Artieres

We propose a non-linear graphical model for structured prediction. It combines the power of deep neural networks to extract high level features with the graphical framework of Markov networks, yielding a powerful and scalable probabilistic model that we apply to signal labeling tasks.

FP9 Structured Prediction Cascades
D. Weiss and B. Taskar

Structured prediction tasks pose a fundamental trade-off between the need for model complexity to increase predictive power and the limited computational resources for inference in the exponentially-sized output spaces such models require. We formulate and develop structured prediction cascades: a sequence of increasingly complex models that progressively filter the space of possible outputs. We represent an exponentially large set of filtered outputs using max marginals and propose a novel convex loss function that balances filtering error with filtering efficiency. We provide generalization bounds for these loss functions and evaluate our approach on handwriting recognition and part-of-speech tagging. We find that the learned cascades are capable of reducing the complexity of inference by up to five orders of magnitude, enabling the use of models which incorporate higher order features and yield higher accuracy.

FP10 Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity
S. Kakade, O. Shamir, K. Sindharan and A. Tewari

The versatility of exponential families, along with their attendant convexity properties, make them a popular and effective statistical model. A central issue is learning these models in high-dimensions when the optimal parameter vector is sparse. This work characterizes a certain strong convexity property of general exponential families, which allows their generalization ability to be quantified. In particular, we show how this property can be used to analyze generic exponential families under L1 regularization.

FP11 Feature Selection using Multiple Streams
P. Dhillon, D. Foster and L. Ungar

Feature selection for supervised learning can be greatly improved by making use of the fact that features often come in classes. For example, in gene expression data, the genes which serve as features may be divided into classes based on their membership in gene families or pathways. When labeling words with senses for word sense disambiguation, features fall into classes including adjacent words, their parts of speech, and the topic and venue of the document the word is in. We present a streamwise feature selection method that allows dynamic generation and selection of features, while taking advantage of the different feature classes, and the fact that they are of different sizes and have different (but unknown) fractions of good features. Experimental results show that our approach provides significant improvement in performance and is computationally less expensive than comparable “batch" methods that do not take advantage of the feature classes and expect all features to be known in advance.

FP12 Fast Active-set-type Algorithms for L1-regularized Linear Regression
J. Kim and H. Park

In this paper, we investigate new active-set-type methods for l1-regularized linear regression that overcome some difficulties of existing active set methods. By showing a relationship between l1-regularized linear regression and the linear complementarity problem with bounds, we present a fast active-set-type method, called block principal pivoting. This method accelerates computation by allowing exchanges of several variables among working sets. We further provide an improvement of this method, discuss its properties, and also explain a connection to the structure learning of Gaussian graphical models. Experimental comparisons on synthetic and real data sets show that the proposed method is significantly faster than existing active set methods and competitive against recently developed iterative methods.

FP13 Nonparametric prior for adaptive sparsity
V. Raykar and L. Zhao

For high-dimensional problems various parametric priors have been proposed to promote sparse solutions. While parametric priors has shown considerable success they are not very robust in adapting to varying degrees of sparsity. In this work we propose a discrete mixture prior which is partially nonparametric. The right structure for the prior and the amount of sparsity is estimated directly from the data. Our experiments show that the proposed prior adapts to sparsity much better than its parametric counterparts. We apply the proposed method to classification of high dimensional microarray datasets.

FP14 Elliptical slice sampling
I. Murray, R. Adams and D. MacKay

Many probabilistic models introduce strong dependencies between variables using a latent multivariate Gaussian distribution or a Gaussian process. We present a new Markov chain Monte Carlo algorithm for performing inference in models with multivariate Gaussian priors. Its key properties are: 1) it has simple, generic code applicable to many models, 2) it has no free parameters, 3) it works well for a variety of Gaussian process based models. These properties make our method ideal for use while model building, removing the need to spend time deriving and tuning updates for more complex algorithms.

FP15 Multi-Task Learning using Generalized t Process
Y. Zhang and D. Yeung

Multi-task learning seeks to improve the generalization performance of a learning task with the help of other related learning tasks. Among the multi-task learning methods proposed thus far, Bonilla et al.'s method provides a novel multi-task extension of Gaussian process (GP) by using a task covariance matrix to model the relationships between tasks. However, learning the task covariance matrix directly has both computational and representational drawbacks. In this paper, we propose a Bayesian extension by modeling the task covariance matrix as a random matrix with an inverse-Wishart prior and integrating it out to achieve Bayesian model averaging. To make the computation feasible, we first give an alternative weight-space view of Bonilla et al.'s multi-task GP model and then integrate out the task covariance matrix in the model, leading to a multi-task generalized t process (MTGTP). For the likelihood, we use a generalized t noise model which, together with the generalized t process prior, brings about the robustness advantage as well as an analytical form for the marginal likelihood. In order to specify the inverse-Wishart prior, we use the maximum mean discrepancy (MMD) statistic to estimate the parameter matrix of the inverse-Wishart prior. Moreover, we investigate some theoretical properties of MTGTP, such as its asymptotic analysis and learning curve. Comparative experimental studies on two common multi-task learning applications show very promising results.

FP16 Efficient Multioutput Gaussian Processes through Variational Inducing Kernels
M. A. Álvarez, D. Luengo, M. Titsias and N. Lawrence

Interest in multioutput kernel methods is increasing, whether under the guise of multitask learning, multisensor networks or structured output data. From the Gaussian process perspective a multioutput Mercer kernel is a covariance function over correlated output functions. One way to construct such kernels is based on convolution processes (CP). A key problem for this approach is efficient inference. Alvarez and Lawrence recently presented a sparse approximation for CPs that enabled efficient inference. In this paper, we extend this work in two directions: we introduce the concept of variational inducing functions to handle potential non-smooth functions involved in the kernel CP construction and we consider an alternative approach to approximate inference based on variational methods, extending the work by Titsias (2009) to the multiple output case. We demonstrate our approaches on prediction of school marks, compiler performance and financial time series.

FP17 Gaussian processes with monotonicity information
J. Riihimäki and A. Vehtari

A method for using monotonicity information in multivariate Gaussian process regression and classification is proposed. Monotonicity information is introduced with virtual derivative observations, and the resulting posterior is approximated with expectation propagation. Behaviour of the method is illustrated with artificial regression examples, and the method is used in a real world health care classification problem to include monotonicity information with respect to one of the covariates.

FP18 On the Convergence Properties of Contrastive Divergence
I. Sutskever and T. Tieleman

Contrastive Divergence (CD) is a popular method for estimating the parameters of Markov Random Fields (MRFs) by rapidly approximating an intractable term in the gradient of the log probability. Despite CD's empirical success, little is known about its theoretical convergence properties. In this paper, we analyze the CD-1 update rule for Restricted Boltzmann Machines (RBMs) with binary variables. We show that this update is not the gradient of any function, and construct a counterintuitive "regularization function" that causes CD learning to cycle indefinitely. Nonetheless, we show that the regularized CD update has a fixed point for a large class of regularization functions using Brower's fixed point theorem.

FP19 Tempered Markov Chain Monte Carlo for training of Restricted Boltzmann Machines
G. Desjardins, A. Courville, Y. Bengio, P. Vincent and O. Delalleau

Alternating Gibbs sampling is the most common scheme used for sampling from Restricted Boltzmann Machines (RBM), a crucial component in deep architectures such as Deep Belief Networks. However, we find that it often does a very poor job of rendering the diversity of modes captured by the trained model. We suspect that this hinders the advantage that could in principle be brought by training algorithms relying on Gibbs sampling for uncovering spurious modes, such as the Persistent Contrastive Divergence algorithm. To alleviate this problem, we explore the use of tempered Markov Chain Monte-Carlo for sampling in RBMs. We find both through visualization of samples and measures of likelihood on a toy dataset that it helps both sampling and learning.

FP20 Learning with Blocks: Composite Likelihood and Contrastive Divergence
A. Asuncion, Q. Liu, A. Ihler and P. Smyth

Composite likelihood methods provide a wide spectrum of computationally efficient techniques for statistical tasks such as parameter estimation and model selection. In this paper, we present a formal connection between the optimization of composite likelihoods and the well-known contrastive divergence algorithm. In particular, we show that composite likelihoods can be stochastically optimized by performing a variant of contrastive divergence with random-scan blocked Gibbs sampling. By using higher-order composite likelihoods, our proposed learning framework makes it possible to trade off computation time for increased accuracy. Furthermore, one can choose composite likelihood blocks that match the model's dependence structure, making the optimization of higher-order composite likelihoods computationally efficient. We empirically analyze the performance of blocked contrastive divergence on various models, including visible Boltzmann machines, conditional random fields, and exponential random graph models, and we demonstrate that using higher-order blocks improves both the accuracy of parameter estimates and the rate of convergence.

FP21 Polynomial-Time Exact Inference in NP-Hard Binary MRFs via Reweighted Perfect Matching
N. Schraudolph

We develop a new form of reweighting (Wainwright et al., 2005) to leverage the relationship between Ising spin glasses and perfect matchings into a novel technique for the exact computation of MAP states in hitherto intractable binary Markov random fields. Our method solves an n by n lattice with external field and random couplings much faster, and for larger n, than the best competing algorithms. It empirically scales as $O(n^3)$ even though this problem is NP-hard and non-approximable in polynomial time. We discuss limitations of our current implementation and propose ways to overcome them.

FP22 HOP-MAP: Efficient Message Passing with High Order Potentials
D. Tarlow, I. Givoni and R. Zemel

There is a growing interest in building probabilistic models with high order potentials (HOPs), or interactions, among discrete variables. Message passing inference in such models generally takes time exponential in the size of the interaction, but in some cases maximum a posteriori (MAP) inference can be carried out efficiently. We build upon such results, introducing two new classes, including composite HOPs that allow us to flexibly combine tractable HOPs using simple logical switching rules. We present efficient message update algorithms for the new HOPs, and we improve upon the efficiency of message updates for a general class of existing HOPs. Importantly, we present both new and existing HOPs in a common representation; performing inference with any combination of these HOPs requires no change of representations or new derivations.

FP23 Why are DBNs sparse?
S. Chatterjee and S. Russell

Real stochastic processes operate in continuous time and can be modeled by sets of stochastic differential equations. On the other hand, several popular model families, including hidden Markov models and dynamic Bayesian networks (DBNs), use discrete time steps. This paper explores methods for converting DBNs with infinitesimal time steps into DBNs with finite time steps, to enable efficient simulation and filtering over long periods. An exact conversion—summing out all intervening time slices between two steps—results in a completely connected DBN, yet nearly all human-constructed DBNs are sparse. We show how this sparsity arises from well-founded approximations resulting from differences among the natural time scales of the variables in the DBN. We define an automated procedure for constructing a provably accurate, approximate DBN model for any desired time step. We illustrate the method by generating a series of approximations to a simple pH model for the human body, demonstrating speedups of several orders of magnitude compared to the original model.

FP24 Graphical Gaussian modelling of multivariate time series with latent variables
M. Eichler

In time series analysis, inference about cause-effect relationships among multiple times series is commonly based on the concept of Granger causality, which exploits temporal structure to achieve causal ordering of dependent variables. One major problem in the application of Granger causality for the identification of causal relationships is the possible presence of latent variables that affect the measured components and thus lead to so-called spurious causalities. In this paper, we describe a new graphical approach for modelling the dependence structure of multivariate stationary time series that are affected by latent variables. To this end, we introduce dynamic maximal ancestral graphs (dMAGs), in which each time series is represented by a single vertex. For Gaussian processes, this approach leads to vector autoregressive models with errors that are not independent but correlated according to the dashed edges in the graph. We discuss identifiability of the parameters and show that these models can be viewed as graphical ARMA models that satisfy the Granger causality restrictions encoded by the associated dynamic maximal ancestral graph.

FP25 Approximation of hidden Markov models by mixtures of experts with application to particle filtering
J. Olsson and J. Ströjby

Selecting conveniently the proposal kernel and the adjustment multiplier weights of the auxiliary particle filter may increase significantly the accuracy and computational efficiency of the method. However, in practice the optimal proposal kernel and multiplier weights are seldom known. In this paper we present a simulation-based method for constructing offline an approximation of these quantities that makes the filter close to fully adapted at a reasonable computational cost. The approximation is constructed as a mixture of experts optimised through an efficient stochastic approximation algorithm. The method is illustrated on two simulated examples.

FP26 Learning Nonlinear Dynamic Models from Non-sequenced Data
T. Huang, L. Song and J. Schneider

Virtually all methods of learning dynamic systems from data start from the same basic assumption: the learning algorithm will be given a sequence, or trajectory, of data generated from the dynamic system. We consider the case where the data is not sequenced. The training data points come from the system's operation but with no temporal ordering. The data are simply drawn as individual disconnected points. While making this assumption may seem absurd at first glance, many scientific modeling tasks have exactly this property. Previous work proposed methods for learning linear, discrete time models under these assumptions by optimizing approximate likelihood functions. In this paper, we extend those methods to nonlinear models using kernel methods. We go on to propose a new approach to solving the problem that focuses on achieving temporal smoothness in the learned dynamics. The result is a convex criterion that can be easily optimized and often outperforms the earlier methods. We test these methods on several synthetic data sets including one generated from the Lorenz attractor.

FP27 Nonparametric Bayesian Matrix Factorization by Power-EP
N. Ding, Y. Qi, R. Xiang, I. Molloy and N. Li

Many real-world applications can be modeled by matrix factorization. By approximating an observed data matrix as the product of two latent matrices, matrix factorization can reveal hidden structures embedded in data. A common challenge to use matrix factorization is determining the dimensionality of the latent matrices from data. Indian Buffet Processes (IBPs) enable us to apply the nonparametric Bayesian machinery to address this challenge. However, it remains a difficult task to learn nonparametric Bayesian matrix factorization models. In this paper, we propose a novel variational Bayesian method based on new equivalence classes of infinite matrices for learning these models. Furthermore, inspired by the success of nonnegative matrix factorization on many learning problems, we impose nonnegativity constraints on the latent matrices and mix variational inference with expectation propagation. This mixed inference method is unified in a power expectation propagation framework. Experimental results on image decomposition demonstrate the superior computational efficiency and the higher prediction accuracy of our methods compared to alternative Monte Carlo and variational inference methods for IBP models. We also apply the new methods to collaborative filtering and role mining and show the improved predictive performance over other matrix factorization methods.

FP28 Collaborative Filtering on a Budget
A. Karatzoglou, A. Smola and M. Weimer

Matrix factorization is a successful technique for building collaborative filtering systems. While it works well on a large range of problems, it is also known for requiring significant amounts of storage for each user or item to be added to the database. This is a problem whenever the collaborative filtering task is larger than the medium-sized Netflix Prize data. In this paper, we propose a new model for representing and compressing matrix factors via hashing. This allows for essentially unbounded storage (at a graceful storage / performance trade-off) for users and items to be represented in a pre-defined memory footprint. It allows us to scale recommender systems to very large numbers of users or conversely, obtain very good performance even for tiny models (e.g. 400kB of data suffice for a representation of the EachMovie problem). We provide both experimental results and approximation bounds for our compressed representation and we show how this approach can be extended to multipartite problems.

FP29 Collaborative Filtering via Rating Concentration
B. Huang and T. Jebara

While most popular collaborative filtering methods use low-rank matrix factorization and parametric density assumptions, this article proposes an approach based on distribution-free concentration inequalities. Using agnostic hierarchical sampling assumptions, functions of observed ratings are provably close to their expectations over query ratings, on average. A joint probability distribution over queries of interest is estimated using maximum entropy regularization. The distribution resides in a convex hull of allowable candidate distributions which satisfy concentration inequalities that stem from the sampling assumptions. The method accurately estimates rating distributions on synthetic and real data and is competitive with low rank and parametric methods which make more aggressive assumptions about the problem.

FP30 Descent Methods for Tuning Parameter Refinement
A. Lorbert and P. Ramadge

This paper addresses multidimensional tuning parameter selection in the context of "train-validate-test" and K-fold cross validation. A coarse grid search over tuning parameter space is used to initialize a descent method which then jointly optimizes over variables and tuning parameters. We study four regularized regression methods and develop the update equations for the corresponding descent algorithms. Experiments on both simulated and real-world datasets show that the method results in significant tuning parameter refinement.

FP31 Learning Policy Improvements with Path Integrals
E. Theodorou, J. Buchli and S. Schaal

With the goal to generate more scalable algorithms with higher efficiency and fewer open parameters, reinforcement learning (RL) has recently moved towards combining classical techniques from optimal control and dynamic programming with modern learning techniques from statistical estimation theory. In this vein, this paper suggests to use the framework of stochastic optimal control with path integrals to derive a novel approach to RL with parametrized policies. While solidly grounded in value function estimation and optimal control based on the stochastic Hamilton-Jacobi-Bellman (HJB) equations, policy improvements can be transformed into an approximation problem of a path integral which has no open parameters other than the exploration noise. The resulting algorithm can be conceived of as model-based, semi-model-based, or even model free, depending on how the learning problem is structured. Our new algorithm demonstrates interesting similarities with previous RL research in the framework of probability matching and provides intuition why the slightly heuristically motivated probability matching approach can actually perform well. Empirical evaluations demonstrate significant performance improvements over gradient-based policy learning and scalability to high-dimensional control problems. We believe that Policy Improvement with Path Integrals PI^2 offers currently one of the most efficient, numerically robust, and easy to implement algorithms for RL based on trajectory roll-outs.

FP32 Efficient Reductions for Imitation Learning
S. Ross and D. Bagnell

Imitation Learning, while applied successfully on many large real-world problems, is typically addressed as a standard supervised learning problem, where it is assumed the training and testing data are i.i.d.. This is not true in imitation learning as the learned policy influences the future test inputs (states) upon which it will be tested. We show that this leads to compounding errors and a regret bound that grows quadratically in the time horizon of the task. We propose two alternative algorithms for imitation learning where training occurs over several episodes of interaction. These two approaches share in common that the learner's policy is slowly modified from executing the expert's policy to the learned policy. We show that this leads to stronger performance guarantees and demonstrate the improved performance on two challenging problems: training a learner to play 1) a 3D racing game (Super Tux Kart) and 2) Mario Bros.; given input images from the games and corresponding actions taken by a human expert and near-optimal planner respectively.

FP33 A Markov-Chain Monte Carlo Approach to Simultaneous Localization and Mapping
P. Torma, A. György and C. Szepesvári

A Markov-Chain Monte Carlo based algorithm is provided to solve the simultaneous localization and mapping (SLAM) problem with general dynamical and observation models under open-loop control and provided that the map-representation is finite dimensional. To our knowledge this is the first provably consistent yet (close-to) practical solution to this problem. The superiority of our algorithm over alternative SLAM algorithms is demonstrated in a difficult loop closing situation.

FP34 Incremental Sparsification for Real-time Online Model Learning
D. Nguyen-Tuong and J. Peters

Online model learning in real-time is required by many applications, for example, robot tracking control. It poses a difficult problem, as fast and incremental online regression with large data sets is the essential component and cannot be realized by straightforward usage of off-the-shelf machine learning methods such as Gaussian process regression or support vector regression. In this paper, we propose a framework for online, incremental sparsification with a fixed budget designed for large scale real-time model learning. The proposed approach combines a sparsification method based on an independency measure with a large scale database. In combination with an incremental learning approach such as sequential support vector regression, we obtain a regression method which is applicable in real-time online learning. It exhibits competitive learning accuracy when compared with standard regression techniques. Implementation on a real robot emphasizes the applicability of the proposed approach in real-time online model learning for real world systems.

Posters from Breaking-News Abstracts

FA1 On predictive distributions in fuzzy Bayesian inference
R. Viertl

FA2 Proximal methods for sparse hierarchical dictionary learning
R. Jenatton, J. Mairal, G. Obozinski and F. Bach

FA3 Estimation in the Burr X parameters and reliability using lower records
A. Bazlizi

FA4 Web scale image annotation: learning to rank with joint word-image embeddings
J. Weston, S. Bengio and N. Usunier

FA5 Unlearning for better mixing
O. Breuleux, Y. Bengio and P. Vincent

FA6 TREERANK: a statistical software for bipartite ranking
N. Basklotis, S. Clemencon, M. Depecker and N. Vayatis

FA7 Online semi-supervised learning on quantized graphs
M. Valko, B. Kveton and L. Huang

FA8 Enhancing the efficieny of spectral clustering with partial supervision
D. Mavroeidis

FA9 Non-parametric change detection using sequential kernel density estimation
G.J. Ross, D.K. Tasoulis and N.M. Adams

FA10 Finding the MAP configuration of a subset of latent variables in a graphical model with discrete variables
G. Teodoru, C. Blundell and M. Sahani

FA11 A statistical test to detect distance concentration from data sets
A. Kaban

Oral sessions: Saturday, May 15th

Invited Talk

08:45 - 09:45 Nonparametric Learning of Functions and Graphs in High Dimensions
John Lafferty

We present recent work on several nonparametric learning problems in the high dimensional setting. In particular, we present theory and methods for estimating sparse regression functions, additive models, and graphical models. For additive models, we present a functional version of methods based on l1 regularization for linear models. For graphical models, we develop methods for estimating the underlying graph based only on observations. One approach is something we call "the nonparanormal," which uses copula methods to transform the variables by nonparametric functions, relaxing the strong distributional assumptions made by the Gaussian graphical model. Another approach is to restrict the family of allowed graphs to spanning forests, enabling the use of fully nonparametric density estimation. All of the approaches are easy to understand, simple to use, theoretically well supported, and effective for modeling of high dimensional data. Joint work with Anupam Gupta, Han Liu, Larry Wasserman, and Min Xu.

Kernel Methods

09:45 - 10:10 Nonlinear functional regression: a functional RKHS approach
H. Kadri

This paper deals with functional regression, in which the input attributes as well as the response are functions. To deal with this problem, we develop a functional reproducing kernel Hilbert space approach; here, a kernel is an operator acting on a function and yielding a function. We demonstrate basic properties of these functional RKHS, as well as a representer theorem for this setting; we investigate the construction of kernels; we provide some experimental insight.

10:10 - 10:35 On the relation between universality, characteristic kernels and RKHS embedding of measures
B. Sriperumbudur, K. Fukumizu and G. Lanckreit

Universal kernels have been shown to play an important role in the achievability of the Bayes risk by many kernel-based algorithms that include binary classification, regression, etc. In this paper, we propose a notion of universality that generalizes the notions introduced by Steinwart and Micchelli et al. and study the necessary and sufficient conditions for a kernel to be universal. We show that all these notions of universality are closely linked to the injective embedding of a certain class of Borel measures into a reproducing kernel Hilbert space (RKHS). By exploiting this relation between universality and the embedding of Borel measures into an RKHS, we establish the relation between universal and characteristic kernels. The latter have been proposed in the context of the RKHS embedding of probability measures, used in statistical applications like homogeneity testing, independence testing, etc.

Graphical Models and Causal Inference

17:00 - 17:25 On combining graph-based variance reduction schemes
V. Gogate and R. Dechter

In this paper, we consider two variance reduction schemes that exploit the structure of the primal graph of the graphical model: Rao-Blackwellised w-cutset sampling and AND/OR sampling. We show that the two schemes are orthogonal and can be combined to further reduce the variance. Our combination yields a new family of estimators which trade time and space with variance. We demonstrate experimentally that the new estimators are superior, often yielding an order of magnitude improvement over previous schemes on several benchmarks.

17:25 - 17:50 Convex structure learning in log-linear models beyond pairwise potentials
M. Schmidt and K. Murphy

Previous work has examined structure learning in log-linear models with L1-regularization, largely focusing on the case of pairwise potentials. In this work we consider the case of models with potentials of arbitrary order, but that satisfy a hierarchical constraint. We enforce the hierarchical constraint using group L1-regularization with overlapping groups, and an active set method that enforces hierarchical inclusion allows us to tractably consider the exponential number of higher-order potentials. We use a spectral projected gradient method as a sub-routine for solving the overlapping group L1-regularization problem, and make use of a sparse version of Dykstra's algorithm to compute the projection. Our experiments indicate that this model gives equal or better test set likelihood compared to previous models.

17:50 - 18:15 Modeling annotator expertise: learning when everybody knows a bit of something
R. Rosales, Y. Yan, G. Fung and J. Dy

Supervised learning from multiple labeling sources is an increasingly important problem in machine learning and data mining. This paper develops a probabilistic approach to this problem when annotators may be unreliable (labels are noisy), but also their expertise varies depending on the data they observe (annotators may have knowledge about different parts of the input space). That is, an annotator may not be consistently accurate (or inaccurate) across the task domain. The presented approach produces classification and annotator models that allow us to provide estimates of the true labels and annotator variable expertise. We provide an analysis of the proposed model under various scenarios and show experimentally that annotator expertise can indeed vary in real tasks and that the presented approach provides clear advantages over previously introduced multi-annotator methods, which only consider general annotator characteristics.

Low-rank Methods and Information Retrieval

18:45 - 19:10 Fluid dynamics models for low rank discriminant analysis
Y.-K. Noh, B.-T. Zhang and D. Lee

We consider the problem of reducing the dimensionality of labeled data for classification. Unfortunately, the optimal approach of finding the low-dimensional projection with minimal Bayes classification error is intractable, so most standard algorithms optimize a tractable heuristic function in the projected subspace. Here, we investigate a physics-based model where we consider the labeled data as interacting fluid distributions. We derive the forces arising in the fluids from information theoretic potential functions, and consider appropriate low rank constraints on the resulting acceleration and velocity flow fields. We show how to apply the Gauss principle of least constraint in fluids to obtain tractable solutions for low rank projections. Our fluid dynamic approach is demonstrated to better approximate the Bayes optimal solution on Gaussian systems, including infinite dimensional Gaussian processes.

19:10 - 19:35 Reduced-rank hidden Markov models
S. Siddiqi, B. Boots and G. Gordon

Hsu et al.(2009) recently proposed an efficient, accurate spectral learning algorithm for Hidden Markov Models (HMMs). In this paper we relax their assumptions and prove a tighter finite-sample error bound for the case of Reduced-Rank HMMs, i.e., HMMs with low-rank transition matrices. Since rank-k RR-HMMs are a larger class of models than k-state HMMs while being equally efficient to work with, this relaxation greatly increases the learning algorithm's scope. In addition, we generalize the algorithm and bounds to models where multiple observations are needed to disambiguate state, and to models that emit multivariate real-valued observations. Finally we prove consistency for learning Predictive State Representations, an even larger class of models. Experiments on synthetic data and a toy video, as well as on difficult robot vision data, yield accurate models that compare favorably with alternatives in simulation quality and prediction accuracy.

19:35 - 20:00 Half transductive ranking
B. Bai, J. Weston, D. Grangier, R. Collobert, C. Cortes and M. Mohri

We study the standard retrieval task of ranking a fixed set of items given a previously unseen query and pose it as the half transductive ranking problem. The task is transductive as the set of items is fixed. Transductive representations (where the vector representation of each example is learned) allow the generation of highly nonlinear embeddings that capture object relationships without relying on a specific choice of features, and require only relatively simple optimization. Unfortunately, they have no direct out-of-sample extension. Inductive approaches on the other hand allow for the representation of unknown queries. We describe algorithms for this setting which have the advantages of both transductive and inductive approaches, and can be applied in unsupervised (either reconstruction-based or graph-based) and supervised ranking setups. We show empirically that our methods give strong performance on all three tasks.

Poster session III: Saturday, May 15th (10:35 - 13:00)

Contributed Posters

SP1 Reducing Label Complexity by Learning From Bags
S. Sabato, N. Srebro and N. Tishby

We consider a supervised learning setting in which the main cost of learning is the number of training labels and one can obtain a single label for a bag of examples, indicating only if a positive example exists in the bag, as in Multi-Instance Learning. We thus propose to create a training sample of bags, and to use the obtained labels to learn to classify individual examples. We provide a theoretical analysis showing how to select the bag size as a function of the problem parameters, and prove that if the original labels are distributed unevenly, the number of required labels drops considerably when learning from bags. We demonstrate that finding a low-error separating hyperplane from bags is feasible in this setting using a simple iterative procedure similar to latent SVM. Experiments on synthetic and real data sets demonstrate the success of the approach.

SP2 Conditional Density Estimation via Least-Squares Density Ratio Estimation
M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya and D. Okanohara

Estimating the conditional mean of an input-output relation is the goal of regression. However, regression analysis is not sufficiently informative if the conditional distribution has multi-modality, is highly asymmetric, or contains heteroscedastic noise. In such scenarios, estimating the conditional distribution itself would be more useful. In this paper, we propose a novel method of conditional density estimation that is suitable for multi-dimensional continuous variables. The basic idea of the proposed method is to express the conditional density in terms of the density ratio and the ratio is directly estimated without going through density estimation. Experiments using benchmark and robot transition datasets illustrate the usefulness of the proposed approach.

SP3 Convexity of Proper Composite Binary Losses
M. Reid and R. Williamson

A composite loss assigns a penalty to a real-valued prediction by associating the prediction with a probability via a link function then applying a class probability estimation (CPE) loss. If the risk for a composite loss is always minimised by predicting the value associated with the true class probability the composite loss is proper. We provide a novel, explicit and complete characterisation of the convexity of any proper composite loss in terms of its link and its “weight function” associated with its proper CPE loss.

SP4 Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries
S. Guo and S. Sanner

Preference elicitation (PE) is an important component of interactive decision support systems that aim to make optimal recommendations to users by actively querying their preferences. In this paper, we outline five principles important for PE in real-world problems: (1) real-time, (2) multiattribute, (3) low cognitive load, (4) robust to noise, and (5) scalable. In light of these requirements, we introduce an approximate PE framework based on TrueSkill for performing efficient closed-form Bayesian updates and query selection for a multiattribute utility belief state — a novel PE approach that naturally facilitates the efficient evaluation of value of information (VOI) heuristics for use in query selection strategies. Our best VOI query strategy satisfies all five principles (in contrast to related work) and performs on par with the most accurate (and often computationally intensive) algorithms on experiments with synthetic and real-world datasets.

SP5 Unsupervised Aggregation for Classification Problems with Large Numbers of Categories
I. Titov, A. Klementiev, K. Small and D. Roth

Classification problems with a very large or unbounded set of output categories are common in many areas such as natural language and image processing. In order to improve accuracy on these tasks, it is natural for a decision-maker to combine predictions from various sources. However, supervised data needed to fit an aggregation model is often difficult to obtain, especially if needed for multiple domains. Therefore, we propose a generative model for unsupervised aggregation which exploits the agreement signal to estimate the expertise of individual judges. Due to the large output space size, this aggregation model cannot encode expertise of constituent judges with respect to every category for all problems. Consequently, we extend it by incorporating the notion of category types to account for variability of the judge expertise depending on the type. The viability of our approach is demonstrated both on synthetic experiments and on a practical task of syntactic parser aggregation.

SP6 Contextual Multi-Armed Bandits
T. Lu, D. Pal and M. Pal

We study contextual multi-armed bandit problems where the context comes from a metric space and the payoff satisfies a Lipschitz condition with respect to the metric. Abstractly, a contextual multi-armed bandit problem models a situation where, in a sequence of independent trials, an online algorithm chooses, based on a given context (side information), an action from a set of possible actions so as to maximize the total payoff of the chosen actions. The payoff depends on both the action chosen and the context. In contrast, context-free multi-armed bandit problems, a focus of much previous research, model situations where no side information is available and the payoff depends only on the action chosen. Our problem is motivated by sponsored web search, where the task is to display ads to a user of an Internet search engine based on her search query so as to maximize the click-through rate (CTR) of the ads displayed. We cast this problem as a contextual multi-armed bandit problem where queries and ads form metric spaces and the payoff function is Lipschitz with respect to both the metrics. For any $\epsilon > 0$ we present an algorithm with regret $O(T^{\frac{a+b+1}{a+b+2}+\epsilon})$ where a, b are the covering dimensions of the query space and the ad space respectively. We prove a lower bound $\Omega(T^{{\frac{\tilde{a}+\tilde{b}+1}{\tilde{a}+\tilde{b}+2}}^{\epsilon}})$ for the regret of any algorithm where $\tilde{a}, \tilde{b}$ are packing dimensions of the query spaces and the ad space respectively. For finite spaces or convex bounded subsets of Euclidean spaces, this gives an almost matching upper and lower bound.

SP7 A Potential-based Framework for Online Multi-class Learning with Partial Feedback
S. Wang, R. Jin and H. Valizadegan

We study the problem of online multi-class learning with partial feedback: in each trial of online learning, instead of providing the true class label for a given instance, the oracle will only reveal to the learner if the predicted class label is correct. We present a general framework for online multi-class learning with partial feedback that adapts the potential-based gradient descent approaches. The generality of the proposed framework is verified by the fact that Banditron is indeed a special case of our work if the potential function is set to be the squared L_2 norm of the weight vector. We propose an exponential gradient algorithm for online multi-class learning with partial feedback. Compared to the Banditron algorithm, the exponential gradient algorithm is advantageous in that its mistake bound is independent from the dimension of data, making it suitable for classifying high dimensional data. Our empirical study with four data sets show that the proposed algorithm for online learning with partial feedback is more effective than the Banditron algorithm.

SP8 Regret Bounds for Gaussian Process Bandit Problems
S. Grünewälder, J. Audibert, M. Opper and J. Shawe-Taylor

Bandit algorithms are concerned with trading exploration with exploitation where a number of options are available but we can only learn their quality by experimenting with them. We consider the scenario in which the reward distribution for arms is modeled by a Gaussian process and there is no noise in the observed reward.

Our main result is to bound the regret experienced by algorithms relative to the a posteriori optimal strategy of playing the best arm throughout based on benign assumptions about the covariance function defining the Gaussian process. We further complement these upper bounds with corresponding lower bounds for particular covariance functions demonstrating that in general there is at most a logarithmic looseness in our upper bounds.

SP9 Active Sequential Learning with Tactile Feedback
H. Saal, J. Ting and S. Vijayakumar

We consider the problem of tactile discrimination, with the goal of estimating an underlying state parameter in a sequential setting. If the data is continuous and high-dimensional, collecting enough representative data samples becomes difficult. We present a framework that uses active learning to help with the sequential gathering of data samples, using information-theoretic criteria to find optimal actions at each time step. We consider two approaches to recursively update the state parameter belief: an analytical Gaussian approximation and a Monte Carlo sampling method. We show how both active frameworks improve convergence, demonstrating results on a real robotic hand-arm system that estimates the viscosity of liquids from tactile feedback data.

SP10 A highly efficient blocked Gibbs sampler reconstruction of multidimensional NMR spectra
J. Won Yoon, S. Wilson and K. Hun Mok

Projection Reconstruction Nuclear Magnetic Resonance (PR-NMR) is a new technique to generate multi-dimensional NMR spectra, which have discrete features that are relatively sparsely distributed in space. A small number of projections from lower dimensional NMR spectra are used to reconstruct the multi-dimensional NMR spectra. We propose an efficient algorithm which employs a blocked Gibbs sampler to accurately reconstruct NMR spectra. This statistical method generates samples in Bayesian scheme. Our proposed algorithm is tested on a set of six projections derived from the three-dimensional 700 MHz HNCO spectrum of HasA, a 187-residue heme binding protein.

SP11 Optimal Allocation Strategies for the Dark Pool Problem
A. Agarwal, P. Bartlett and M. Dama

We study the problem of allocating stocks to dark pools. We propose and analyze an optimal approach for allocations, if continuous-valued allocations are allowed. We also propose a modification for the case when only integer-valued allocations are possible. We extend the previous work on this problem by Ganchev et al (UAI 2009) to adversarial scenarios, while also improving over their results in the iid setup. The resulting algorithms are efficient, and are tested on extensive simulations under stochastic and adversarial inputs. Our work also has consequences for other perishable inventory control problems, extending their analyses to adversarial models too.

SP12 Multitask Learning for Brain-Computer Interfaces
M. Alamgir, M. Grosse-Wentrup and Y. Altun

Brain-computer interfaces (BCIs) are limited in their applicability in everyday settings by the current necessity to record subject-specific calibration data prior to actual use of the BCI for communication. In this paper, we utilize the framework of multitask learning to construct a BCI that can be used without any subject-specific calibration process. We discuss how this out-of-the-box BCI can be further improved in a computationally efficient manner as subject-specific data becomes available. The feasibility of the approach is demonstrated on two sets of experimental EEG data recorded during a standard two-class motor imagery paradigm from a total of 19 healthy subjects. Specifically, we show that satisfactory classification results can be achieved with zero training data, and combining prior recordings with subject-specific calibration data substantially outperforms using subject-specific data only. Our results further show that transfer between recordings under slightly different experimental setups is feasible.

SP13 An Alternative Prior Process for Nonparametric Bayesian Clustering
H. Wallach, S. Jensen, L. Dicker and K. Heller

Prior distributions play a crucial role in Bayesian approaches to clustering. Two commonly-used prior distributions are the Dirichlet and Pitman-Yor processes. In this paper, we investigate the predictive probabilities that underlie these processes, and the implicit "rich-get-richer" characteristic of the resulting partitions. We explore an alternative prior for nonparametric Bayesian clustering, the uniform process, for applications where the "rich-get-richer" property is undesirable. We also explore the cost of this new process: partitions are no longer exchangeable with respect to the ordering of variables. We present new asymptotic and simulation-based results for the clustering characteristics of the uniform process and compare these with known results for the Dirichlet and Pitman-Yor processes. Finally, we compare performance on a real document clustering task, demonstrating the practical advantage of the uniform process despite its lack of exchangeability over orderings.

SP14 Matrix-Variate Dirichlet Process Mixture Models
Z. Zhang, G. Dai and M. Jordan

We are concerned with a multivariate response regression problem where the interest is in considering correlations both across response variates and across response samples. In this paper we develop a new Bayesian nonparametric model for such a setting based on Dirichlet process priors. Building on an additive kernel model, we allow each sample to have its own regression matrix. Although this overcomplete representation could in principle suffer from severe overfitting problems, we are able to provide effective control over the model via a matrix-variate Dirichlet process prior on the regression matrices. Our model is able to share statistical strength among regression matrices due to the clustering property of the Dirichlet process. We make use of a Markov chain Monte Carlo algorithm for inference and prediction. Compared with other Bayesian kernel models, our model has advantages in both computational and statistical efficiency.

SP15 Dependent Indian Buffet Processes
S. Williamson, P. Orbanz and Z. Ghahramani

Latent variable models represent hidden structure in observational data.To account for the distribution of the observational data changing over time, space or some other covariate, we need generalizations of latent variable models that explicitly capture this dependency on the covariate. A variety of such generalizations has been proposed for latent variable models based on the Dirichlet process. We address dependency on covariates in binary latent feature models, by introducing a dependent Indian buffet process. The model generates, for each value of the covariate, a binary random matrix with an unbounded number of columns. Evolution of the binary matrices over the covariate set is controlled by a hierarchical Gaussian process model. The choice of covariance functions controls the dependence structure and exchangeability properties of the model. We derive a Markov Chain Monte Carlo sampling algorithm for Bayesian inference, and provide experiments on both synthetic and real-world data. The experimental results show that explicit modeling of dependencies significantly improves accuracy of predictions.

SP16 Posterior distributions are computable from predictive distributions
C. Freer and D. Roy

As we devise more complicated prior distributions, will inference algorithms keep up? We highlight a negative result in computable probability theory by Ackerman, Freer, and Roy (2010) that shows that there exist computable priors with noncomputable posteriors. In addition to providing a brief survey of computable probability theory geared towards the A.I. and statistics community, we give a new result characterizing when conditioning is computable in the setting of exchangeable sequences, and provide a computational perspective on work by Orbanz (2010) on conjugate nonparametric models. In particular, using a computable extension of de Finetti's theorem (Freer and Roy 2009), we describe how to transform a posterior predictive rule for generating an exchangeable sequence into an algorithm for computing the posterior distribution of the directing random measure.

SP17 A generalization of the Multiple-try Metropolis algorithm for Bayesian estimation and model selection
S. Pandolfi, F. Bartolucci and N. Friel

We propose a generalization of the Multiple-try Metropolis (MTM) algorithm of Liu et al. (2000), which is based on drawing several proposals at each step and randomly choosing one of them on the basis of weights that may be arbitrary chosen. In particular, for Bayesian estimation we also introduce a method based on weights depending on a quadratic approximation of the posterior distribution. The resulting algorithm cannot be reformulated as an MTM algorithm and leads to a comparable gain of efficiency with a lower computational effort. We also outline the extension of the proposed strategy, and then of the MTM strategy, to Bayesian model selection, casting it in a Reversible Jump framework. The approach is illustrated by real examples.

SP18 Sequential Monte Carlo Samplers for Dirichlet Process Mixtures
Y. Ulker, B. Günsel and T. Cemgil

In this paper, we develop a novel online algorithm based on the Sequential Monte Carlo(SMC) samplers framework for posterior inference in Dirichlet Process Mixtures (DPM). Our method generalizes many sequential importance sampling approaches. It provides a computationally efficient improvement to particle filtering that is less prone to getting stuck in isolated modes. The proposed method is a particular SMC sampler that enables us to design sophisticated clustering update schemes, such as updating past trajectories of the particles in light of recent observations, and still ensures convergence to the true DPM target distribution asymptotically. Performance has been evaluated in a Bayesian Infinite Gaussian mixture density estimation problem and it is shown that the proposed algorithm outperforms conventional Monte Carlo approaches in terms of estimation variance and average log-marginal likelihood.

SP19 Learning Bayesian Network Structure using LP Relaxations
T. Jaakkola, D. Sontag, A. Globerson and M. Meila

We propose to solve the combinatorial problem of finding the highest scoring Bayesian network structure from data. This structure learning problem can be viewed as an inference problem where the variables specify the choice of parents for each node in the graph. The key combinatorial difficulty arises from the global constraint that the graph structure has to be acyclic. We cast the structure learning problem as a linear program over the polytope defined by valid acyclic structures. In relaxing this problem, we maintain an outer bound approximation to the polytope and iteratively tighten it by searching over a new class of valid constraints. If an integral solution is found, it is guaranteed to be the optimal Bayesian network. When the relaxation is not tight, the fast dual algorithms we develop remain useful in combination with a branch and bound method. Empirical results suggest that the method is competitive or faster than alternative exact methods based on dynamic programming.

SP20 Bayesian structure discovery in Bayesian networks with less space
P. Parviainen and M. Koivisto

Current exact algorithms for score-based structure discovery in Bayesian networks on n nodes run in time and space within a polynomial factor of 2^n. For practical use, the space requirement is the bottleneck, which motivates trading space against time. Here, previous results on finding an optimal network structure in less space are extended in two directions. First, we consider the problem of computing the posterior probability of a given arc set. Second, we operate with the general partial order framework and its specialization to bucket orders, introduced recently for related permutation problems. The main technical contribution is the development of a fast algorithm for a novel zeta transform variant, which may be of independent interest.

SP21 Inference of Sparse Networks with Unobserved Variables. Application to Gene Regulatory Networks
N. Slavov

Networks are becoming a unifying framework for modeling complex systems and network inference problems are frequently encountered in many fields. Here, I develop and apply a generative approach to network inference (RCweb) for the case when the network is sparse and the latent (not observed) variables affect the observed ones. From all possible factor analysis (FA) decompositions explaining the variance in the data, RCweb selects the FA decomposition that is consistent with a sparse underlying network. The sparsity constraint is imposed by a novel method that significantly outperforms (in terms of accuracy, robustness to noise, complexity scaling and computational efficiency) methods using l1 norm relaxation such as K-SVD and l1-based sparse principle component analysis (PCA). Results from simulated models demonstrate that RCweb recovers exactly the model structures for sparsity as low (as non-sparse) as 50% and with ratio of unobserved to observed variables as high as 2. RCweb is robust to noise, with gradual decrease in the parameter ranges as the noise level increases.

SP22 Simple Exponential Family PCA
J. Li and D. Tao

Bayesian principal component analysis (BPCA), a probabilistic reformulation of PCA with Bayesian model selection, is a systematic approach to determining the number of essential principal components (PCs) for data representation. However, it assumes that data are Gaussian distributed and thus it cannot handle all types of practical observations, e.g. integers and binary values. In this paper, we propose simple exponential family PCA (SePCA), a generalised family of probabilistic principal component analysers. SePCA employs exponential family distributions to handle general types of observations. By using Bayesian inference, SePCA also automatically discovers the number of essential PCs. We discuss techniques for fitting the model, develop the corresponding mixture model, and show the effectiveness of the model based on experiments.

SP23 Locally Linear Denoising on Image Manifolds
D. Gong, F. Sha and G. Medioni

We study the problem of image denoising where images are assumed to be samples from low dimensional (sub)manifolds. We propose the algorithm of locally linear denoising. The algorithm approximates manifolds with locally linear patches by constructing nearest neighbor graphs. Each image is then locally denoised within its neighborhoods. A global optimal denoising result is then identified by aligning those local estimates. The algorithm has a closed-form solution that is efficient to compute. We evaluated and compared the algorithm to alternative methods on two image data sets. We demonstrated the effectiveness of the proposed algorithm, which yields visually appealing denoising results, incurs smaller reconstruction errors and results in lower error rates when the denoised data are used in supervised learning tasks.

SP24 Supervised Dimension Reduction Using Bayesian Mixture Modeling
K. Mao, F. Liang and S. Mukherjee

We develop a Bayesian framework for supervised dimension reduction using a flexible nonparametric Bayesian mixture modeling approach. Our method retrieves the dimension reduction or d.r. subspace by utilizing a dependent Dirichlet process that allows for natural clustering for the data in terms of both the response and predictor variables. Formal probabilistic models with likelihoods and priors are given and efficient posterior sampling of the d.r. subspace can be obtained by a Gibbs sampler. As the posterior draws are linear subspaces which are points on a Grassmann manifold, we output the posterior mean d.r. subspace with respect to geodesics on the Grassmannian. The utility of our approach is illustrated on a set of simulated and real examples. Some Key Words: supervised dimension reduction, inverse regression, Dirichlet process, factor models, Grassman manifold.

SP25 Hartigan's Method: k-means Clustering without Voronoi
M. Telgarsky and A. Vattani

Hartigan's method for k-means clustering is the following greedy heuristic: select a point, and optimally reassign it. This paper develops two other formulations of the heuristic, one leading to a number of consistency properties, the other showing that the data partition is always quite separated from the induced Voronoi partition. A characterization of the volume of this separation is provided. Empirical tests verify not only good optimization performance relative to Lloyd's method, but also good running time.

SP26 Towards Understanding Situated Natural Language
A. Bordes, N. Usunier, R. Collobert and J. Weston

We present a general framework and learning algorithm for the task of concept labeling: each word in a given sentence has to be tagged with the unique physical entity (e.g. person, object or location) or abstract concept it refers to. Our method allows both world knowledge and linguistic information to be used during learning and prediction. We show experimentally that we can learn to use world knowledge to resolve ambiguities in language, such as word senses or reference resolution, without the use of handcrafted rules or features.

SP27 Discriminative Topic Segmentation of Text and Speech
M. Mohri, P. Moreno and E. Weinstein

We explore automated discovery of topically-coherent segments in speech or text sequences. We give two new discriminative topic segmentation algorithms which employ a new measure of text similarity based on word co-occurrence. Both algorithms function by finding extrema in the similarity signal over the text, with the latter algorithm using a compact support-vector based description of a window of text or speech observations in word similarity space to overcome noise introduced by speech recognition errors and off-topic content. In experiments over speech and text news streams, we show that these algorithms outperform previous methods. We observe that topic segmentation of speech recognizer output is a more difficult problem than that of text streams; however, we demonstrate that by using a lattice of competing hypotheses rather than just the one-best hypothesis as input to the segmentation algorithm, the performance of the algorithm can be improved.

SP28 State-Space Inference and Learning with Gaussian Processes
R. Turner, M. Deisenroth and C. Rasmussen

State-space inference and learning with Gaussian processes (GPs) is an unsolved problem. We propose a new, general methodology for inference and learning in nonlinear state-space models that are described probabilistically by non-parametric GP models. We apply the expectation maximization algorithm to iterate between inference in the latent state-space and learning the parameters of the underlying GP dynamics model.

SP29 Model-Free Monte Carlo-like Policy Evaluation
R. Fonteneau, S. Murphy, L. Wehenkel and D. Ernst

We propose an algorithm for estimating the finite-horizon expected return of a closed loop control policy from an a priori given (off-policy) sample of one-step transitions. It averages cumulated rewards along a set of “broken trajectories” made of one-step transitions selected from the sample on the basis of the control policy. Under some Lipschitz continuity assumptions on the system dynamics, reward function and control policy, we provide bounds on the bias and variance of the estimator that depend only on the Lipschitz constants, on the number of broken trajectories used in the estimator, and on the sparsity of the sample of one-step transitions.

SP30 Variational methods for Reinforcement Learning
T. Furmston and D. Barber

We consider reinforcement learning as solving a Markov decision process with unknown transition distribution. Based on interaction with the environment, an estimate of the transition matrix is obtained from which the optimal decision policy is formed. The classical maximum likelihood point estimate of the transition model does not reflect the uncertainty in the estimate of the transition model and the resulting policies may consequently lack a sufficient degree of exploration. We consider a Bayesian alternative that maintains a distribution over the transition so that the resulting policy takes into account the limited experience of the environment. The resulting algorithm is formally intractable and we discuss two approximate solution methods, Variational Bayes and Expectation Propagation.

SP31 Efficient Learning of Deep Boltzmann Machines
R. Salakhutdinov and H. Larochelle

We present a new approximate inference algorithm for Deep Boltzmann Machines (DBM's), a generative model with many layers of hidden variables. The algorithm learns a separate "recognition" model that is used to quickly initialize, in a single bottom-up pass, the values of the latent variables in all hidden layers. We show that using such a recognition model, followed by a combined top-down and bottom-up pass, it is possible to efficiently learn a good generative model of high-dimensional highly-structured sensory input. We show that the additional computations required by incorporating a top-down feedback plays a critical role in the performance of a DBM, both as a generative and discriminative model. Moreover, inference is only at most three times slower compared to the approximate inference in a Deep Belief Network (DBN), making large-scale learning of DBM's practical. Finally, we demonstrate that the DBM's trained using the proposed approximate inference algorithm perform well compared to DBN's and SVM's on the MNIST handwritten digit, OCR English letters, and NORB visual object recognition tasks.

SP32 Inductive Principles for Restricted Boltzmann Machine Learning
B. Marlin, K. Swersky, B. Chen and N. de Freitas

Recent research has seen the proposal of several new inductive principles designed specifically to avoid the problems associated with maximum likelihood learning in models with intractable partition functions. In this paper, we study learning methods for binary restricted Boltzmann machines (RBMs) based on ratio matching and generalized score matching. We compare these new RBM learning methods to a range of existing learning methods including stochastic maximum likelihood, contrastive divergence, and pseudo-likelihood. We perform an extensive empirical evaluation across multiple tasks and data sets.

SP33 Why Does Unsupervised Pre-training Help Deep Learning?
D. Erhan, A. Courville, Y. Bengio and P. Vincent

Much recent research has been devoted to learning algorithms for deep architectures such as Deep Belief Networks and stacks of auto-encoder variants with impressive results being obtained in several areas, mostly on vision and language datasets. The best results obtained on supervised learning tasks often involve an unsupervised learning component, usually in an unsupervised pre-training phase. The main question investigated here is the following: why does unsupervised pre-training work so well? Through extensive experimentation, we explore several possible explanations discussed in the literature including its action as a regularizer (Erhan et al. 2009) and as an aid to optimization (Bengio et al. 2007). Our results build on the work of Erhan et al. 2009, showing that unsupervised pre-training appears to play predominantly a regularization role in subsequent supervised training. However our results in an online setting, with a virtually unlimited data stream, point to a somewhat more nuanced interpretation of the roles of optimization and regularization in the unsupervised pre-training effect.

SP34 Parallelizable Sampling of Markov Random Fields
J. Martens and I. Sutskever

Markov Random Fields (MRFs) are an important class of probabilistic models which are used for density estimation, classification, denoising, and for constructing Deep Belief Networks. Every application of an MRF requires addressing its inference problem, which can be done using deterministic inference methods or using stochastic Markov Chain Monte Carlo methods. In this paper we introduce a new Markov Chain transition operator that updates all the variables of a pairwise MRF in parallel by using auxiliary Gaussian variables. The proposed MCMC operator is extremely simple to implement and to parallelize. This is achieved by a formal equivalence result between arbitrary pairwise MRFs and a particular type of Restricted Boltzmann Machine. This result also implies that the later can be learned in place of the former without any loss of modeling power, a possibility we explore in experiments.

Posters from Breaking-News Abstracts

SA1 Nested sampling and the foundations of computational inference
J. Skilling

SA2 Analysis of finite sample effects in compressed Fisher's LDA
R.J. Durrant and A. Kaban

SA3 Optimal rates for conjugate gradient regularization
G. Blanchard and N. Kramer

SA4 How to save feature extraction time for fast and robust classification?
J. Louradour and C. Kermorvant

SA5 Greedy learning of binary latent trees
S. Narmeling and C.K.I. Williams

SA6 Bayesian spatial models for hospital recruitment using integrated nested Laplace approximation
M. Musio, E.-A. Sauleau and V. Mameli

SA7 Estimating the contribution of non-genetic factors to gene expression using GP-LVM
N. Fusi and N. Lawrence

SA8 Drifting linear dynamics
T. Raiko, A. Ilin, N. Korsakova, E. Oja and L. Valpola

SA9 Likelihood unimodality of a state-space model with point process observations
K. Yuan and M. Niranjan

SA10 Learning the iHMM through iterative map-reduce
S. Bratieres, J. van Gael, A. Vlachos and Z. Ghahramani

SA11 Classification of functional data: a weighted distance approach
A.M. Alonso, D. Casado and J.J. Romo