Mathematical Programming in Machine Learning and Data Mining
January 14th-January 19th, 2007

MEALS

*Breakfast (Buffet): 7:00–9:00 am, Donald Cameron Hall, Monday–Friday
*Lunch (Buffet): 11:30 am–1:30 pm, Donald Cameron Hall, Monday–Friday
*Dinner (Buffet): 5:30–7:30 pm, Donald Cameron Hall, Sunday–Thursday
Coffee Breaks: As per daily schedule, 2nd floor lounge, Corbett Hall
***Please remember to scan your meal card at the host/hostess station in the dining room for each meal.**

MEETING ROOMS

All lectures will be held in Max Bell 159 (Max Bell Building accessible by bridge on 2nd floor of Corbett Hall). Hours: 6 am–12 midnight. LCD projector, overhead projectors and blackboards are available for presentations. *Please note that the meeting space designated for BIRS is the lower level of Max Bell, Rooms 155–159. Please respect that all other space has been contracted to other Banff Centre guests, including any Food and Beverage in those areas.*

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk - Professional Development Centre - open 24 hours)
	Lecture rooms available after 16:00 (if desired)
17:30–19:30	Buffet Dinner, Donald Cameron Hall
20:00	Informal gathering in 2nd floor lounge, Corbett Hall
	Beverages and small assortment of snacks available on a cash honour-system.

Monday

7:00–8:45	Breakfast
8:45–9:00	Introduction and Welcome to BIRS by BIRS Station Manager, Max Bell 159
9:00–10:00	Michael J. Todd "On minimum-volume ellipsoids: From John and Kiefer-Wolfowitz to Khachiyan and Nesterov-Nemirovski"
10:00–10:30	Zhaosong Lu "Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration complexity for cone programming"
10:30–11:00	Coffee Break, 2nd floor lounge, Corbett Hall
11:00–12:00	Alexandre d'Aspermont, "Semidefinite Optimization with Applications in Sparse Multivariate Statistics"
12:00–13:00	Lunch
13:00–14:00	Tony Jebara, "Semidefinite Programming for Classification and Dimensionality Reduction"
14:00–15:00	Gert Lanckriet, TBA
15:00–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:30	Nathan Srebro, TBA
16:30–17:30	Kilian Weinberger, TBA
17:30–18:00	Personal time/discussions
18:00–19:30	Dinner

Tuesday

7:00–8:00	Breakfast
8:00–9:00	Wotao Yin, "Various optimization approaches for imaging problems"
9:00–10:00	Thorsten Joachims, "Large-Margin Training for Predicting Structured Outputs"
10:00–10:30	
	Group Photo; meet on the front steps of Corbett Hall (to take place directly after the last lecture of the morning) (can be scheduled for a different time or day if required)
10:30–11:00	Coffee Break, 2nd floor lounge, Corbett Hall
11:00–12:00	Personal time
12:00–13:00	Lunch
13:00–14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00–15:00	Personal time
15:00–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:30	Chris Burges, "Learning to Rank"
16:30–18:00	Steve Wright's talk and a panel discussion on new directions and new classes of optimization methods in machine learning" Panel coordinators: Steve Wright, Grace Wahba, Dale Shuurmans, Don Goldfarb
18:00–19:30	Dinner
19:30–20:30	John Langford, "The reductive relationships between learning problems" and a continued discussion

Wednesday

7:00–8:00	Breakfast
8:00–9:00	Sam Roweis, "Visualizing Pairwise Similarity via Semidefinite Programming"
9:00–9:30	Francis Bach, "Low-rank matrix factorization with attributes"
10:00–10:30	Joydeep Ghosh, "Locating a Few Good Clusters: A Tale of Two Viewpoints"
10:00–10:30	Coffee Break, 2nd floor lounge, Corbett Hall
10:30–11:00	Takashi Tsuchiya, "A recursive recomputation approach for smoothing in nonlinear and Bayesian state-space modeling"
11:00–11:30	Ted Wild, "Nonlinear Knowledge in Kernel Machines"
11:30–13:30	Lunch
	Free Afternoon
17:30–19:30	Dinner

Thursday

7:00–9:00	Breakfast
9:00–10:00	Inderjit Dhillon, "Machine Learning with Bregman Divergences"
10:00–10:30	Grace Wahba, "Selection of high order patterns in demographic and genetic data"
10:30–11:00	Coffee Break, 2nd floor lounge, Corbett Hall
11:00–12:00	Jong-Shi Pang, "Bilevel Optimization and Machine Learning"
12:00–13:00	Lunch
13:00–14:00	Shai Ben David, TBA
14:00–15:00	Panos Pardalos, "Biclustering in Data Mining"
15:00–15:30	Coffee Break, 2nd floor lounge, Corbett Hall
15:30–16:00	Yasemin Altun, "Regularization in Learning to Predict Structured Objects"
16:00–17:00	Marina Meila, "The stability of a good clustering"
17:00–18:00	Wine service and closing remarks, 2nd floor lounge, Corbett Hall
18:00–19:30	Dinner

Friday

7:00–9:00	Breakfast
9:00	Informal Discussions, as many participants must catch early flights
	Coffee Break, 2nd floor lounge, Corbett Hall - 10 am
11:30–13:30	Lunch

Checkout by 12 noon.

** 5-day workshops are welcome to use the BIRS facilities (2nd Floor Lounge, Max Bell Meeting Rooms, Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. **

Mathematical Programming in Machine Learning and Data Mining
January 14th-19th, 2007

ABSTRACTS
(in alphabetic order by speaker surname)

Speaker: **Alexandre d'Aspermont** (Princeton)
Title: *Semidefinite Optimization with Applications in Sparse Multivariate Statistics*
Abstract:

We use recently developed first order methods for semidefinite programming to solve convex relaxations of combinatorial problems arising in sparse multivariate statistics. We discuss in detail applications to sparse principal component analysis, sparse covariance selection and sparse nonnegative matrix factorization.

Speaker: **Yasemin Altun** (TTI, Chigaco)
Title: *Regularization in Learning to Predict Structured Objects*
Abstract:

Predicting objects with complex structure is ubiquitous in many application areas. Recent work on machine learning focused on devising different loss functions and algorithms for structured output prediction. Another important component of learning is regularization and it has not been explored in structured output prediction problems till now. However, the complex structure of the outputs results in learning with features with dramatically different properties, which in turn can require different regularizations. Convex analysis tools provide the connection between regularization and approximate moment matching constraints. Motivated with these theoretical results, we explore various regularization schemes in learning to predict structured outputs, in particular hierarchical classification and label sequence learning.

Speaker: **Francis Bach** (Ecole des Mines de Paris)
Title: *Low-rank matrix factorization with attributes*
Abstract:

We develop a new collaborative filtering (CF) method that combines both previously known users' preferences, i.e. standard CF, as well as product/user attributes, i.e. classical function approximation, to predict a given user's interest in a particular product. Our method is a generalized low rank matrix completion problem, where we learn a function whose inputs are pairs of vectors – the standard low rank matrix completion problem being a special case where the inputs to the function are the row and column indices of the matrix. We solve this generalized matrix completion problem using tensor product kernels for which we also formally generalize standard kernel properties. Benchmark experiments on movie ratings show the advantages of our generalized matrix completion method over the standard matrix completion one with no information about movies or people, as well as over standard multi-task or single task learning methods.

Speaker: **Chris Burges** (Microsoft Research)
Title: *Learning to Rank*
Abstract:

The problem of ranking occurs in many guises. The field of Information Retrieval is largely dependent on ranking: there the problem is, given a query, to sort a (sometimes huge) database of documents in order of relevance. Recommender systems also often need to rank: given a set of movies or songs that some collaborative filtering algorithm has decided you would probably enjoy, which ones should be at the top of the list? Ranking has been less studied in the machine learning community than classification, but the two are also closely related: for a binary classifier, the area under the ROC curve (the curve of true

positives versus false positives) is equal to a simple ranking statistic. In this talk I will give an overview of the problem from the point of view of the needs of a large, commercial search engine. I will describe some recent approaches to solving the ranking problem. Considering this problem highlights a serious problem in machine learning that is rarely addressed: the mismatch between the cost functions we optimize, and the ones we actually care about. I will also describe recent work that is aimed at addressing this "optimization / target cost mismatch" problem.

Speaker: **Joydeep Ghosh** (UT Austin)
Title: *Locating a Few Good Clusters: A Tale of Two Viewpoints*
Abstract:

Many applications involve discovering a small number of dense or cohesive clusters in the data while ignoring the bulk of the data. We will discuss two broad approaches to this problem: (a) a generative approach where one determines and fits a suitable probablistic model to the data, and (b) a non-parametric approach inspired by Wishart's remarkable but obscure mode analysis work from 1968. The pros and cons of the two approaches will be illustrated using results from both artificial and gene expression data analysis.

Speaker: **Tony Jebara** (Columbia University)
Title: *Semidefinite Programming for Classification and Dimensionality Reduction*
Abstract:

We propose semidefinite programming (SDP) to improve the support vector machine (SVM) linear classifier by exploiting tighter Vapnik-Chervonenkis (VC) bounds based on an ellipsoidal gap-tolerant classification model. SDPs are used to modify the regularization criterion for a linear classifier which improves its accuracy dramatically without making any additional assumptions on the binary classification problem. A bounding minimum volume ellipsoid is estimated via SDP on the data and used to redefine the margin in an SVM. The technique is fully kernelizable and therefore accommodates nonlinear classification as well. Tighter VC generalization bounds can also be estimated numerically using an iterated variant of SDP.

In addition, a similar iterated variant of SDP is used to improve dimensionality reduction by directly optimizing the eigen-gap. This method is reminiscent of semidefinite embedding which reduces dimensionality of the data by maximizing the trace of a matrix (the sum of the eigenvalues). Our novel method gives rise to a more general linear function of the eigenvalues in the SDP which is handled iteratively by interleaving the SDP with eigen-decomposition. In some cases, only global minima exist for these general linear functions of eigenvalues. Experiments reveal that this is a competitive method for visualizing high dimensional data.

Speaker: **Thorsten Joachims** (Cornell University)
Title: *Large-Margin Training for Predicting Structured Outputs*
Abstract:

Over the last decade, much of the research on discriminative learning has focused on problems like classification and regression, where the prediction is a single univariate variable. But what if we need to predict complex objects like trees, orderings, or alignments? Such problems arise, for example, when a natural language parser needs to predict the correct parse tree for a given sentence, when one needs to optimize a multivariate performance measure like the F1-score, or when predicting the alignment between two proteins.

This talk discusses how these complex and structured prediction problems can be formulated as convex programs. In particular, it presents a support vector approach that generalizes conventional classification SVMs to a large range of structured outputs and multivariate loss functions. The resulting optimization problems are convex quadratic, but have an exponential (or infinite) number of constraints. To solves the training problems efficiently, the talk explores a cutting-plane algorithm. The algorithm is implemented in the SVM-Struct software and empirical results will be given for several examples.

Speaker: **Zhaosong Lu, Ph.D.** (Simon Fraser University)
Title: *Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration complexity for cone programming*
Abstract:

In this talk we consider the general cone programming problem, and propose primal-dual convex (smooth and/or nonsmooth) minimization reformulations for it. We then discuss first-order methods suitable for solving these reformulations, namely, Nesterov's optimal method, Nesterov's smooth approximation scheme, and Nemirovski's prox-method, and propose a variant of Nesterov's optimal method which has outperformed the latter one in our computational experiments. We also derive iteration-complexity bounds for these first-order methods applied to the proposed primal-dual reformulations of the cone programming problem. The performance of these methods is then compared using a set of randomly generated linear programming and semidefinite programming (SDP) instances. We also compare the approach based on the variant of Nesterov's optimal method with the low-rank method proposed by Burer and Monteiro for solving a set of randomly generated SDP instances.

Speaker: **Marina Meila** (University of Washington)
Title: *The stability of a good clustering*
Abstract: If we have found a "good" clustering C of data set X, can we prove that C is not far from the (unknown) best clustering C* of this data set? Perhaps surprisingly, the answer to this question is sometimes yes. We can show bounds on the distance(C, C*) for two clustering criteria: the Normalized Cut and the squared distance cost of K-means clustering. These bounds exist in the case when the data X admits a "good" clustering for the given cost.

Speaker: **Panos Pardalos** (University of Florida)
Title: *Biclustering in Data Mining*
Abstract:

Biclustering consists of simultaneous partitioning of the set of samples and the set of their attributes (features) into subsets (classes). Samples and features classified together are supposed to have a high relevance to each other. We review the most widely used and successful biclustering techniques and their related applications from a theoretical viewpoint emphasizing mathematical concepts that can be met in existing biclustering techniques. Then we define the notion of consistency for biclustering using interrelation between centroids of sample and feature classes. We have shown that consistent biclustering implies separability of the classes by convex cones. While earlier works on biclustering concentrated on unsupervised learning and did not consider employing a training set, whose classification is given, our model represents supervised biclustering, whose consistency is achieved by feature selection. It involves the solution of a fractional 0-1 programming problem. Encouraging computational results on microarray data mining problems are reported.

Speaker: **Sam Roweis** (U.Toronto)
Title: *Visualizing Pairwise Similarity via Semidefinite Programming*
Abstract:

Binary pairwise similarity data is available in many domains where quantifying the similarity/difference between objects is extremely difficult or impossible. Nonetheless, it is often desirable to obtain insight into such data by associating each object (record) with a point in some abstract feature space – for visualization purposes this space is often two or three dimensional. We present an algorithm for visualizing such similarity data, which delivers an embedding of each object such that similar objects are always closer in the embedding space than dissimilar ones. Many such mappings may exist, and our method selects amongst them the one in which the mean distance between embedded points is as large as possible. This has the effect of stretching the mapping and, interestingly, favoring embeddings with low effective dimensionality.

We study both the parametric and non-parametric variants of the problem, showing that they both result in convex Semidefinite Programs (SDP). In the non-parametric version, input points may be mapped

to any point in space, whereas the parametric version assumes that the mapping is given by some function (e.g. a linear or kernel mapping) of the input. This allows us to generalize the embedding to points not used in the training procedure.

Speaker: **Michael J. Todd** (Cornell University)
Title: *On minimum-volume ellipsoids: From John and Kiefer-Wolfowitz to Khachiyan and Nesterov-Nemirovski*
Abstract:

The problem of finding the minimum-volume ellipsoid containing a set in R^n has arisen in contexts from optimization to statistics and data analysis over the last sixty years. We describe some of these settings and algorithms old and new for solving the problem.

Speaker: **Takashi Tsuchiya** (The Institute of Statistical Mathematics)
Title: *A recursive recomputation approach for smoothing in nonlinear and Bayesian state-space modeling*
Abstract:

Smoothing in the state-space model is to compute the state distribution under the condition that the whole observation is given. Smoothing has a number of applications in signal processing, statistics, pattern recognition, machine learning, bioinformatics, speech recognition, target tracking etc. One of the major difficulties in numerical smoothing for nonlinear models is that it requires storage proportional to the length T of the time series to store the whole filtering distribution.

We develop a generic implementation scheme for numerical smoothing which drastically reduces the space complexity from O(T) to O(log T) by using the recursive recomputation technique developed for automatic differntiation by Griewank. We accomplish this reduction by computing the whole filtering distribution O(log T) times instead of once.

The Japanese stock market price time series data with T=956 is taken up as an instance to demonstrate advantage of the proposed scheme. The particle filter is implemented with the new scheme to smooth the whole interval estimating the change of volatility. The number of particles is 3,000,000, and the whole interval is smoothed with 5.3GB storage, accomplishing saving of storage by a factor of 1/20. The ordinary implementation would require 106GB and hence would be intractable.

Speaker: **Grace Wahba** (U. of Wisconsin)
Title: *Selection of high order patterns in demographic and genetic data*
Abstract:

We give a description of the recent LASSO-Patternsearch algorithm and its application to the 'mining' of higher order patterns/interactions in demographic and genetic data via statistical model building methods. Recent applications include discovery of multiple interacting risk factors for an eye disease, and a high degree of accuracy in separating cases from controls based on patterns of SNP data. A new conputational algorithm capable of selecting from thousands of possible patterns at once is described. Partial results can be found in Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., and Klein, B. " LASSO-Patternsearch Algorithm with Application to Ophthalmalogy Data. " TR 1131, October, 2006, available via the TRLIST on my home page http://www.stat.wisc.edu/ wahba. Various generalizations will be described if time permits.

Speaker: **Ted Wild** (U. of Wisconsin)
Title: *Nonlinear Knowledge in Kernel Machines*
Abstract:

We give a unified presentation of recent work in applying prior knowledge to nonlinear kernel approximation and nonlinear kernel classification. In both approaches, prior knowledge over general nonlinear sets is incorporated into nonlinear kernel approximation or classification problems as linear constraints in a linear program. The key tool in this incorporation is a theorem of the alternative for convex functions that converts nonlinear prior knowledge implications into linear inequalities without the need to kernelize these

implications. Effectiveness of the proposed approximation formulation is demonstrated on two synthetic examples as well as an important lymph node metastasis prediction problem arising in breast cancer prognosis. Effectiveness of the proposed classification formulation is demonstrated on three publicly available datasets, including a breast cancer prognosis dataset. All these problems exhibit marked improvements upon the introduction of prior knowledge of nonlinear kernel approximation and classification approaches that do not utilize such knowledge.