Mathematical Programming in Machine Learning and Data Mining

Katya Scheinberg (IBM TJ Watson Research Center),
Jiming Peng (University of Illinois at Urbana-Champaign),
Tamas Terlaky (McMaster University),
Dale Shuurmans (University of Alberta),
Michael Jordan (UC Berkeley),
Tomaso Poggio (MIT)

January 14–January 19, 2007

1 Overview of the Field

The field of Machine Learning (ML) and Data Mining (DM) is focused around the following problem: Given a data domain D we want to approximate an unknown function $y(x)$ on the given data set $X \subset D$ (for which the values of $y(x)$ may or may not be known) by a function f from a given class $\mathcal{F}$ so that the approximation generalizes in the best possible way on all of the (unseen) data $x \in D$. The approximating function f might take real values, as in the case of regression; binary values, as in the case of classification; or integer values, as in some cases of ranking; or this function might be a mapping between ordered subsets of data points and ordered subsets of real, integer or binary values, as in the case of structured object prediction. The quality of approximation by f can be measured by various objective functions. For instance in the case of support vector machine (SVM)[4] classification the quality of the approximating function is estimated by a weighted sum of a regularization term $h(f)$ and the hinge loss term $\sum_{x \in X} \max\{1-y(x)f(x), 0\}$. Hence, many of the machine learning problems can be posed as an optimization problem where optimization is performed over a given class $\mathcal{F}$ for a chosen objective.

The connection between optimization and machine learning (although always present) became especially evident with the popularity of the SVMs [4], [24], and the kernel methods in general [18]. SVM classification problem is formulated as a convex quadratic program.

$$
(P) \qquad \begin{array}{ll}
\max & -\frac{1}{2}\alpha^T Q\alpha - c\sum_{i=1}^{n}\xi_i \\
\text{s.t.} & -Q\alpha + by + s - \xi = -e, \\
 & 0 \le \alpha \le c,\ s \ge 0,\ \xi \ge 0,
\end{array}
$$

where $\alpha \in \mathbf{R}^n$ is the vector of dual variables, b is the bias (scalar) and s and ξ are the n-dimensional vectors of slack and surplus variables, respectively. y is a vector of labels, e is the vector of all 1's of length n and c is the penalty parameter of the loss function in the objective. Q is the label encoded kernel matrix, i.e. $Q_{ij} = y_i y_j K(x_i, x_j)$, where $K(\cdot,\cdot)$ the the kernel matrix (or function) which implicitely defines the class $\mathcal{F}$.

The problem is always feasible and, in theory, finding the global solution for this problem is easy, that is it can be done in polynomial time. However, many large-scale practical cases proved to be difficult to handle by standard optimization software. This led to a number of special purpose implementations. First implementations were developed by researchers from the ML community [15], [10] and some of the later implementations were proposed by the mathematical programming (MP) community [7], [6], [5], [17].

Success of the convex optimization for SVMs led to the extensive use of convex optimization models and methods for other machine learning problems in the past 6-7 years, such as learning the kernel [12], [3], where optimization is done over the matrices $K(x_i, x_j)$; and computation of the entire regularization path [8], [16], where soltion is found for all possible values of the penalty parameter c. Beyond classification, optimization model have been used in dimensionality reduction [9], [26]; low rank matrix factorization [20], [11], metric learning [29], [25], structured objects prediction [23], [22] and many others. It became apparent that the connection between the two fields can benefit greatly from the collaboration of the researches from both sides. However, the MP community and the ML/DM community are quite disjoint. They do not typically share conferences or publication venues. To remedy this situation in the past several years there have been several occasions when researches from both fields were brought together in a workshop or a conference. There were two workshop organized specifically on optimization in machine learning and data mining. Both were held in 2005, one held in Trunau,Germany and one at McMaster University in Canada. A special topic on machine learning and large scale optimization was published in the Journal of Machine Learning Research in 2006. The potential for collaboration between the fields have been steadily increasing in the last few years.

The purpose of the Banff workshop was to continue and improve upon the effort of bringing together outstanding researches from the fields of mathematical programming, data mining and statistical machine learning to ignite new collaborations and expose each side to the possibilities available in each field. The purpose is to identify the new problems and to match them with potential solution approaches.

2 Recent Developments and Open Problems

The amount of literature in the Machine Learning and Data Mining communities that involves optimization models is very extensive. We do not attempt to present a comprehensive survey of existing and possible topics here. We focus, instead, on the work immediately relevant to the results presented at workshop. This work, in fact, is a representative selection of recent developments and open problems in the field, but it is by no means axhaustive.

2.1 Nonlinear Classification

Nonlinear classification via kernels is the essence of support vector machines. There are still many unresolved questions, perhaps main of them being, how to find a good kernel. This question has been addressed in recent years with limited success by means of convex optimization [12], [3].

The two issues discussed at the workshop were of a different nature, however. It is well known that the use of kernels in SVM is made possible by the use of ℓ_2 regularization term on the model parameters. It is unclear how to extend kernalization for the ℓ_1 regularization, which otherwise may produce better (more sparse) models. Saharon Rosset in his talk discussed the extension of the ℓ_1 regularization case to the infinite dimensional case which allows the use of nonlinear ℓ_1 classification. Ted Wild addressed topic of exploiting prior knowledge when using a kernel.

2.2 Structured objects prediction

Structured objects prediction is a generalization of the standard classification or regression problems. Instead of predicting a label of an object the structured object prediction aim to predict a set of labels of a collection of objects. Structured output classification can be posed similarly to a classification problem but with exponential number of constraints. Due to the very large number of constraints and a somewhat different motivation for structured output prediction the extension of classification approaches in not straightforward. A few talk of the workshop addressed this issue. We list ranking in the category of structured output prediction problems because these problems are related. In both cases the number of constraints is too large to search over exhaustively and in both cases the optimization function (the loss function and the regularization) is more complex to derive than for the classification and regression problems. Of the talks on the topic presented at the conference one talk (by Yasemin Altun) addressed the question of introducing regularization into the structured output prediction problems. Another talk (by Chris Burges) discussed the modeling approaches specifically to the ranking problem and the third talk by Thorsten Joachims addressed an SVM formulation and a fast optimization method for predicting structured objects.

2.3 Dimensionality reduction and metric learning

Semidefinite programming (SDP)[28] was by far the most popular optimization topic of the workshop. There is a large variety of machine learning problems that are being posed as semidefinite programming problems. The main application of semidefinite programming arises in dimensionality reduction. If the data is given to us as a set of points in a space of a large dimension it is important to recognize if in actuality it lies on (or near to) a low dimensional surface. Alternatively, the data can be embedded in a low dimensional space by using a proper distance metric. If the data is represented by a matrix, it may be desirable to extract a low rank matrix factorization, or a low rank matrix completion. All of these setting can be addressed via an SDP formulation. In some case the SDP is a relaxation of the original nonconvex problem. Several talks at the workshop presented SDP models for ML problems.

A fundamental difficulty with the SDP approach is that the resulting SDPs are expensive to solve in practice with conventional methods. Recently new methods for solving large scale SDPs were proposed in the optimization community. These method have inferior performance guarantees, compared to the interior point methods, but they can be much faster in practice for certain classes of SDPs, expecially if an accurate solution is not needed [13],[14]. Fast approximate methods which exploit the structure of the specific formulations were discussed in the talk of d'Aspermont and Weinberger. A special case of solving SDPs - the problem of finding minimum volume enclosing ellipsoid was presented by Todd [2].

2.4 Clustering

Robust clustering that deals with the uncertainty, noise in the data set, which a major concern in cluster analysis. Dr. Ghosh's talk consider a special case of robust clustering where the target is to find interesting patterns in the data set by ignoring a certain number of outliers. New optimization-based approaches have been proposed. This is different from the traditional approach where the outliers were removed first based on statistical measurements.

In many scenarios such as in biological discovery and diagnosis, we need to find not only the patterns in the data set, but also the features that characterize these patterns. This leads to the so-called bi-clustering problem, which has recently become a very hot research topic in the clustering community. Pardalos' talk proposed an fractional program model to attack this problem. He also showed that the problem is NP-hard and suggested some heuristics to solve the new optimization model applied to biological applications.

Meila's talk considered the robustness of the data set, i.e., under what distribution, a 'good' clustering can converge to the best clustering.

Ben-David's talk addressed the complexity issues of clustering.

2.5 Complexity

The issues of empirical and theoretical complexity of many of the machine learning problems are still unresolved. The theoretical complexity bounds for the underlying optimization models are usually known, but they are typically based on the worst case analysis and assume that the goal is to find a fairly accurate solution. It is not always relevant for the ultimate goal of a machine learning problem - good generalization error bounds. Some of these issued were addressed at the workshop. A talk by John Langford (Yahoo! Inc.) focused on reductions of ML problems. The goal is to find relationship between machine learning problems, and construct algorithms which work via reduction. Another talk, by Nathan Srebro addressed the issue of how complexity of a ML problem should depend on the size of the available data set.

3 Presentation Highlights

We will now list some of the presentations arranged by the topics discussed above.

3.1 Nonlinear Classification

Speaker: **Saharon Rosset** (IBM Research)
Title: ℓ_1 *Regularization in Infinite Dimensional Feature Spaces*
Abstract:

In this talk I discuss the problem of fitting ℓ_1 regularized prediction models in infinite (possibly non-countable) dimensional feature spaces. Our main contributions are: a. Deriving a generalization of ℓ_1 regularization based on measures which can be applied in non-countable feature spaces; b. Proving that the sparsity property of ℓ_1 regularization is maintained in infinite dimensions; c. Devising a path-following algorithm that can generate the set of regularized solutions in ”nice” feature spaces; and d. Presenting an example of penalized spline models where this path following algorithm is computationally feasible, and gives encouraging empirical results.

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.

3.2 Structured Objects Prediction

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: **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 a 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: **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.

3.3 Demensionality Reduction and Semidefinite Programming

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: **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: **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: **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: **Kilian Weinberger** (University of Pennsylvania)
Title: *Distance Metric Learning via Semidefinite Programming*
Abstract: Many problems in computer science can be simplified by clever representations of sensory or symbolic input. How to discover such representations automatically, from large amounts of data, remains a fundamental challenge. The goal of metric learning is to derive Euclidean representations of labeled or unlabeled inputs from observed statistical regularities. In this talk I will review two recently proposed algorithms for metric learning. Both algorithms rely on modern tools in convex optimization that are proving increasingly useful in many areas of machine learning. In addition to the two metric learning algorithms, I will propose a novel method [27] to approximate large scale SDPs with Laplacian graph regularization.

3.4 Clustering

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 probabilistic 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: **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.

3.5 Complexity

Speaker: **Nathan Srebro** (IBM Research & TTI-Chicago)
Title: *Computational Complexity and Data Set Size*
Abstract:

In devising methods for optimization problems associated with learning tasks, and in studying the runtime of these methods, we usually think of the runtime as increasing with the data set size. However, from a learning performance perspective, having more data available should not mean we need to spend more time optimizing. At the extreme, we can always ignore some of the data if it makes optimization difficult. But perhaps having more data available can actually allow us to spend less time optimizing?

It appears that such behavior exists in several combinatorial problems such as learning the dependency structure of Markov networks, sparse sensing, and Gaussian-mixture clustering. In these problems there appears to be a phase transition, where learning beneath some data threshold is computationally intractable (although it is statistically possible), but learning with more data becomes computationally easy. This threshold was empirically studied and characterized for the problem of Gaussian-mixture clustering [21].

Can perhaps a more continuous, but similar, effect exist in convex optimization problems such as learning Support Vector Machines (SVMs)? A new stochastic gradient decent approach for SVM learning [19] does in-fact display such behavior: the computation time required to obtain a predictor with some target accuracy decreases, rather than increases, with the amount of available data.

4 Scientific Progress Made

Here we list the impact which the workshop already have had on the work of some of the participants. We are sure that there are other participants whose work also have been or is likely to be affected by the workshop, but we do not have the complete information.

4.1 Don Goldfarb's report

Here is some feedback on the conference. Specifically, here are some of the ways that the conference has impacted my own research.

Jong-Shi Pang's talk on bi-level optimization and machine learning introduced me to new applications of bi-level programming. As these problems can be formulated as math programs with equilibrium (or complementarity) constraints, this is important to me as I have been working on developing algorithms for these problems. In fact I just submitted a paper on this subject to SIOPT.

I found the talk by Inderjit Dhillon on "Machine Learning with Bregman Distances" very useful as I have also used Bregman distances in recent work on image denoising. I expect that I will be able to apply some of my ideas to the topics discussed by Dhillon.

I also found the talks by Kilian Weinberger, Gert Lanckriet, Tony Jabara, Alexandre D'Aspremont and Sam Roweis covering various uses of semidefinite programming in machine learning to be of great interest as this is an area in which I am also working.

4.2 Ted Wild's report

The workshop provided me with an excellent opportunity to interact with other researchers. I particularly enjoyed the talk on the representer theorem for 1-norm support vector machines and the talk on parameter selection, both of which gave me ideas I hope to implement in future research.

One topic that seemed to arise frequently at the workshop was dealing with (dis-)similarity data. Applications and methods for processing such data were discussed. Often, the solution involved learning a low-rank kernel matrix or low-dimensional surface via convex programming.

4.3 Alexandre D'Aspermont's report

Francis Bach and I wrote a paper while we were in Banff, it's under review at ICML. A title and abstract follow:

"Full Regularization Path for Sparse Principal Component Analysis"

Given a sample covariance matrix, we examine the problem of maximizing the variance explained by a particular linear combination of the input variables while constraining the number of nonzero coefficients in this combination. This is known as sparse principal component analysis and has a wide array of applications in machine learning and engineering. We formulate a new semidefinite relaxation to this problem and derive a greedy algorithm that computes a *full set* of good solutions for all numbers of non zero coefficients, with complexity $O(n^3)$, where n is the number of variables. We then use the same relaxation to derive sufficient conditions for global optimality of a solution, which can be tested in $O(n^3)$. We show on toy examples and biological data that our algorithm does provide globally optimal solutions in many cases.

4.4 Nathan Srebro

I can point out several direct benefits of the workshop on my work in the short time since the workshop:

Discussions with Bach and d'Aspremont, including hearing of current work by Bach, that were directly helpful in current work on trace-norm regularization submitted to ICML. This directly lead to the optimization method we are now using, and clarified relationship between different formulations.

Comments and pointers from Zhang on stochastic gradient methods that had large impact on directions of current work on fast SVM optimization also submitted to ICML.

Hearing about recent progress in the optimization literature, most notably about the LR method for SDPs, which directly relates to my work.

Other relevant pointers to specific papers, some of which I have already used directly.

Working with another workshop participant (Rosset) on a paper already accepted for publication.

I except interactions and discussions at the workshop would lead to even more collaborations. This was by far the most productive workshop I've participate in.

4.5 Chris Burges's report

Probably the biggest impact for me was generated by John Langford's talk, which I found very intriguing: I had not heard of these results before, and will now follow their development. I also liked Michael Todd's talk, and the fact that the mathematical programming crowd took as a take-home message that we need faster ways to solve SDPs. It was also very valuable simply to chat with various people, some of whom I knew beforehand, some not. Re. your question re. main methods, etc., for ranking, there is really increasing interest in this, from rather standard methods like neural nets, to treating ranking as a structured learning problem a la Joachims etc., to brand new methods (that I touched upon) to solve problems with non-differentiable costs. I think this is a key open problem in machine learning, that the ML community is only now gradually catching on to: the standard method, of coming up with a smooth cost function that attempts to encapsulate the problem, but whose form is chosen as much for computational tractability as for fidelity to the problem being solved, is very likely leading us astray, and is demonstrably doing so, in the case of information retrieval. I don't think the Mathematical Programming community has this as a hot topic, at the moment, though.

4.6 Tony Jebara's report

Overall, the workshop was excellent, it really helped to talk to a small yet high-expertise crowd who are actively using and developing the tools we are spending time with. Here are some more specific benefits:

We integrated some elements of my presentation as well as concepts from Gert Lanckriet's talk into 2 lectures in an advanced machine learning course (COMS 6998-4 Learning and Empirical Inference) being offered at Columbia by Vladimir Vapnik, Gerald Tesauro, Irina Rish and myself.

We explored some of the matrix factorization ideas in my lab after seeing some related concepts at BIRS.

After a brief conversation with folks at BIRS, we realized that a more general set of spectral functions can be optimized with our method and have an abstract on this at the Learning Workshop in 2007 in Puerto Rico.

Don Goldfarb and Watao Yin briefly discussed possibilities for collaboration, possibly when the term ends and schedules are less hectic.

4.7 Katya Scheinberg's report

The workshop broadened and substatially extended my understanding of the field of machine learning and the current use of optimization.

Two of the participants Wotao Yin and Alexandre D'Aspermont will visit IBM in spring to give presentations and discuss possible research ideas for collaboration with another participant Saharon Rosset and myself.

I was invited to visit Joydeep Ghosh and Indrajit Dhillon at the University of Texas at Austin, where I presented my work on active set methods for support vector machines and where we discussed other optimization approaches for convex QPs arising in nonnegative matrix factorization, which Indrajit is interested in.

The talk by Thorsten Joachims was very interesting for me and I have since read related papers and had a few ideas that I am planning to use in my work.

The talk by Kilian Weinberger also was related to some ideas on metric learning which we were exploring with my colleague at IBM.

4.8 Jiming Peng's report

There were quite a few interesting talks. Steve Wright and Inderjit Dhillon have been in contact and collaboration. Following Dr. Ghosh's talk on robust clustering, where the purpose is to find several well-structured clusters that cover a certain portion of the data set, had a 2-hour meeting with Joydeep, and we discussed how his problem can be modeled as 0-1 conic optimization, a fresh optimization model proposed by myself. We are working on a joint project along this line.

The talks by two students are also very interesting. One is by Kilian Weinberger on dimension reduction where the purpose is to represent high-dimensional data in a lower-dimensional space while reserving certain relationships in the original space, another one is by Ted Wild on incorporating prior knowledge into the SVM approaches for better separation.

Jong Shi Pang's talk starts to set up a bridge between machine learning and bi-level optimization, and it seems to have a lot of potential in the future.

5 Outcome of the Meeting

The meeting was viewed by all of the participants as very successful. Quoting Nathan Srebro " This was by far the most productive workshop I've participate in." The overall quality of the talks was outstanding. The workshop attracted researchers with very closely related interests, yet whose work covers a fairly broad set of machine learning problems. As a result each talk covered a new topic and caused intense and lively discussions. Most of the talk were given one hour time slots, which allowed for many questions during the talks. Some of the talks were thus transformed into informative group discussion and it greatly benefited the understanding by the audience. Such impromptu discussion are only possible in the informal setting of a small workshop such as this.

The schedule contained two long free periods (of five to six hours) which we used for recreational activities such as hiking and skiing and also for uninterrupted collaboration work. There were also 2 presentations/discussion scheduled for after dinner time, which continued on informal basis in the lounge for the remainder of the evenings.

There were 34 participants, including 5 students and 5 women. Along with participants from academia there were participants from IBM, Microsoft, Yahoo! Inc. and TTI Research. Below is the complete list of participant.
Yasemin Altun, Toyota Technological Institute
Serhat Aybat, Columbia University
Francis Bach, Center of Mathematical Morphology
Shai Ben-David, University of Waterloo
Chris Burges, Microsoft Research
Alexandre d'Aspremont, Princeton University
Inderjit Dhillon, University of Texas, Austin
Joydeep Ghosh, University of Texas, Austin
Donald Goldfarb, Columbia University
Tony Jebara, Columbia University
Thorsten Joachims, Cornell University
Gert Lanckriet, University of California, San Diego
John Langford, Yahoo Inc.
Sang Lee, University of Wisconsin-Madison
Marina Meila, University of Washington
Hans Mittelmann, Arizona State University
Jong-Shi Pang, Rensselaer Polytechnic Institute
Panos Pardalos, University of Florida
Jiming Peng, University of Illinois at Urbana-Champaign
Saharon Rosset, IBM Research
Sam Roweis, University of Toronto
Katya Scheinberg, IBM TJ Watson Research Center
Dale Schuurmans, University of Alberta
Nathan Srebro, University of Toronto
Michael Todd, Cornell University
Takashi Tsuchiya, Tokyo Institute of Statistical Mathematics
Grace Wahba, University of Wisconsin, Madison
Kilian Weinberger,University of Pennsylvania
Ted Wild, University of Wisconsin-Madison
Steve Wright, University of Wisconsin
Wotao Yin, Rice University

Tong Zhang, Yahoo! Inc.
Xiaojin (Jerry) Zhu, University of Wisconsin-Madison
Jiaping Zhu, McMaster University

References

[1] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, G. R. G. Lanckriet, A Direct Formulation for Sparse PCA using Semidefinite Programming, *Advances in Neural Information Processing Systems 17* (2004), MIT Press, Cambridge, MA.

[2] S.D. Ahipasaoglu, P. Sun, and M.J. Todd, Linear convergence of a modified Frank-Wolfe algorithm for computing minimum volume enclosing ellipsoids, (2006), Technical Report, Cornell University.

[3] F. Bach, G. R. G. Lanckriet and M. I. Jordan, Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proceedings of the 21st International Conference on Machine Learning* (2004), Banff, Canada, Omnipress.

[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Macines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000

[5] E. M. Gertz and S. J. Wright, Object-Oriented Software for Quadratic Programming,*A*CM Transactions on Mathematical Software **29** (2003), 58–81.

[6] M. C. Ferris and T. S. Munson. Interior point methods for massive support vector machines, *SIAM Journal on Optimization*, **13** (2003), 783–804.

[7] S. Fine and K. Scheinberg, Efficient SVM Training Using Low-Rank Kernel Representations, *Journal of Machine Learning Research*, **2**, 2001, 243–264.

[8] T. Hastie, S. Rosset, R. Tibshirani and J. Zhu, The Entire Regularization Path for the Support Vector Machine, *Journal of Machine Learning Research* **5** (2004) 1391–1415.

[9] P. Shivaswamy and T. Jebara. Ellipsoidal Kernel Machines, *Artificial Intelligence and Statistics, AISTATS* (2007).

[10] , T. Joachims, Making large-scale support vector machine learning practical *Advances in Kernel Methods, Schölkopf, B. and Burges, C. C. and Smola, A. J., eds.* **12** 169–184, MIT Press, 1999.

[11] D. Kim, S. Sra, and I. S. Dhillon, Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem. To appear in *Proceedings of the Sixth SIAM International Conference on Data Mining* (2007).

[12] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, L and M. I. Jordan, Learning the Kernel Matrix with Semidefinite Programming, *Journal of Machine Learning Research* **5** (2004), 27–72.

[13] Zh. Lu, A. S. Nemirovski and R. D. C. Monteiro, Large-Scale Semidefinite Programming via Saddle Point Mirror-Prox Algorithm, *Mathematical Programming* **109**(2) (2007), 211–237.

[14] Yu. Nesterov, Sooth minimization of nonsmooth functions, *Mathematical Programming*, **103** (2005), 127–152.

[15] , J. C. Platt, Fast Training Support Vector Machines Using Sequential Mininal Optimization *Advances in Kernel Methods, Schölkopf, B. and Burges, C. C. and Smola, A. J., eds.* **12** 185–208, MIT Press, 1999.

[16] S. Rosset, J. Zhu and T. Hastie, Boosting as a Regularized Path to A Maximum Margin Classifier, *Journal of Machine Learning Research*, **5** (2004) 941–973.

[17] K. Scheinberg, An Efficient Implementation of an Active Set Method for SVMs, *Journal of Machine Learning Research*, **7**, (2006), 2237–2257.

[18] B. Schlkopf and A. Smola. *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[19] S. Shalev-Shwartz, Y. Singer and N. Srebro, Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *24th International Conference on Machine Learning (ICML)*, (2007).

[20] N. Srebro, J. Rennie and T. Jaakkola, Maximum Margin Matrix Factorization for Collaborative Prediction. In *Advances in Neural Information Processing Systems (17)* (2005).

[21] N. Srebro G. Shakhnarovich and S. Roweis, An Investigation of Computational and Informational Limits in Gaussian Mixture Clustering. In *23rd International Conference on Machine Learning (ICML)* (2006).

[22] B. Taskar, C. Guestrin and D. Koller, Max-Margin Markov Networks. In *Advances in Neural Information Processing Systems (16)* (2004).

[23] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research (JMLR)*, **6** (2005), 1453–1484.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[25] K. Q. Weinberger, J. Blitzer, and L. K. Saul, Distance Metric Learning for Large Margin Nearest Neighbor Classification, In *Advances in Neural Information Processing Systems (18) (Y. Weiss, B. Schoelkopf, and J. Platt eds.)* (2006), MIT Press: Cambridge, MA.

[26] K. Q. Weinberger, F. Sha, and L. K. Saul, Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the Twenty First International Confernence on Machine Learning (ICML-04)* (2004), Banff, Canada.

[27] K. Weinberger and F. Sha and Q. Zhu and L. Saul, Graph Laplacian methods for large-scale semidefinite programming, with an application to sensor localization. In *Advances in Neural Information Processing Systems (19), B. Schölkopf, J. Platt and Thomas Hofmann, eds*, MIT Press, Cambridge, MA, 2007.

[28] H. Wolkowicz, R. Saigal and L. Vandenberghe, Eds., *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*, Kluwer, 2000.

[29] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems (15), (S. Becker, S. Thrun, and K. Obermayer, Eds.)* (2003).