December 13, 2008, 07:30–10:30 and 15:30–18:30 Westin: Nordic **WS17**

Kernel Learning: Automatic Selection of Optimal Kernels

http://www.cs.nyu.edu/learning_kernels

Corinna Cortes corinna@google.com
Google Research New York
Arthur Gretton arthur@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics
Gert Lanckriet gert@ece.ucsd.edu
University of Califorina, San Diego
Mehryar Mohri mohri@cims.nyu.edu
Courant Institute of Mathematical Sciences & Google Research
Afshin Rostamizadeh rostami@cs.nyu.edu
Courant Institute of Mathematical Sciences

Abstract

Kernel methods are widely used to address a variety of learning tasks including classification, regression, ranking, clustering, and dimensionality reduction. The appropriate choice of a kernel is often left to the user. But, poor selections may lead to sub-optimal performance. Furthermore, searching for an appropriate kernel manually may be a time-consuming and imperfect art. Instead, the kernel selection process can be included as part of the overall learning problem. In this way, better performance guarantees can be given and the kernel selection process can be made automatic. In this workshop, we will be concerned with using sampled data to select or learn a kernel function or kernel matrix appropriate for the specific task at hand. We will discuss several scenarios, including classification, regression, and ranking, where the use of kernels is ubiquitous, and different settings including inductive, transductive, or semi-supervised learning. We also invite discussions on the closely related fields of features selection and extraction, and are interested in exploring further the connection with these topics. The goal is to cover all questions related to the problem of learning kernels: different problem formulations, the computational efficiency and accuracy of the algorithms that address these problems and their different strengths and weaknesses, and the theoretical guarantees provided. What is the computational complexity? Does it work in practice? The formulation of some other learning problems, e.g. multi-task learning problems, is often very similar. These problems and their solutions will also be discussed in this workshop.

7:30-8:00	**Invited Speaker: Shai Ben-David** The Sample Complexity of Learning the Kernel
8:00-8:20	**Olivier Chapelle and Alain Rakotomamonjy** Second Order Optimization of Kernel Parameters
8:20-8:50	**Invited Speaker: William Stafford Noble** Multi-Kernel Learning for Biology
8:50-9:20	Poster Session and Discussion
9:20-9:40	**Corinna Cortes, Mehryar Mohri and Afshin Rostamizadeh** Learning Sequence Kernels
9:40-10:00	**Maria-Florina Balcan, Avrim Blum and Nathan Srebro** Learning with Multiple Similarity Functions

10:00-10:30	**Invited Speaker: Andreas Argyriou** Multi-Task Learning via Matrix Regularization
10:30-15:30	Break until afternoon session.
15:30-16:00	**Invited Speaker: Isabelle Guyon** Feature Selection: From Correlation to Causality
16:00-16:20	**Nathan Srebro and Shai Ben-David** Learning Bounds for Support Vector Machines with Learned Kernels
16:20-16:50	**Invited Speaker: Alex Smola** Mixed Norm Kernels, Hyperkenels and Other Variants
16:50-17:20	Poster Session and Discussion
17:20-17:40	**Marius Kloft, Ulf Brefeld, Pavel Laskov and Sören Sonnenburg** Non-sparse Multiple Kernel Learning
17:40-18:00	**Peter Gehler** Infinite Kernel Learning
18:00-18:30	**Invited Speaker: John Shawe-Taylor** Kernel Learning for Novelty Detection
18:30	Closing Remarks

The Sample Complexity of Learning the Kernel
Shai Ben-David, University of Waterloo
The success of kernel based learning algorithms depends upon the suitability of the kernel to the learning task. Ideally, the choice of a kernel should based on prior information of the learner about the task at hand. However, in practice, kernel parameters are being tuned based on available training data. I will discuss the sample complexity overhead associated with such "learning the kernel" scenarios. I will address the setting in which the training data for the kernel selection is target labeled examples, as well as settings in which this training is based on different types of data, such as unlabeled examples and examples labeled by a different (but related) tasks. Part of this work is joint with Nati Srebro.

Second Order Optimization of Kernel Parameters
Olivier Chapelle et al., Yahoo! Research & University Rouen
We investigate the use of second order optimization approaches for solving the multiple kernel learning (MKL) problem. We show that the hessian of the MKL can be computed efficiently and this information can be used to compute a better descent direction than the gradient (used in the state-of-the-art SimpleMKL algorithm). We then empirically show that our new approaches outperforms SimpleMKL in terms of computational efficiency.

Multi-Kernel Learning for Biology
William Stafford Noble, University of Washington
One of the primary tasks facing biologists today is to integrate the different views of molecular biology that are provided by various types of experimental data. In yeast, for example, for a given gene we typically know the protein it encodes, that protein's similarity to other proteins, the mRNA expression levels associated with the given gene under hundreds of experimental conditions, the occurrences of known or inferred transcription factor binding sites in the upstream region of that gene, and the identities of many of the proteins that interact with the given gene's protein product. Each of these distinct data types provides one view of the molecular machinery of the cell.

Kernel methods allow us to represent these heterogeneous data types in a normal form, and to use kernel algebra to reason about more than one type of data simultaneously. Consequently, multi-kernel learning

methods have been applied to a variety of biology applications. In this talk, I will describe several of these applications, outline the lessons we have learned from applying multi-kernel learning methods to real data, and suggest several avenues for future research in this area.

Learning Sequence Kernels

Corinna Cortes et al., Google Research & Courant Institute

Kernel methods are used to tackle a variety of learning tasks including classification, regression, ranking, clustering, and dimensionality reduction. The appropriate choice of a kernel is often left to the user. But, poor selections may lead to a sub-optimal performance. Instead, sample points can be used to learn a kernel function appropriate for the task by selecting one out of a family of kernels determined by the user. This paper considers the problem of *learning sequence kernel functions*, an important problem for applications in computational biology, natural language processing, document classification and other text processing areas. For most kernel-based learning techniques, the kernels selected must be positive definite symmetric, which, for sequence data, are found to be rational kernels. We give a general formulation of the problem of learning rational kernels and prove that a large family of rational kernels can be learned efficiently using a simple quadratic program both in the context of support vector machines and kernel ridge regression. This improves upon previous work that generally results in a more costly semi-definite or quadratically constrained quadratic program. Furthermore, in the specific case of kernel ridge regression, we give an alternative solution for the optimal *kernel matrix*, which in fact coincides with the objective prescribed by kernel alignment techniques.

Learning with Multiple Similarity Functions

Maria-Florina Balcan et al., Microsoft Research & Carnegie Mellon University & Toyota Technological Institute

Kernel functions have become an extremely popular tool in machine learning, with many applications and an attractive theory. There has also been substantial work on learning kernel functions from data [LCBGJ04,SB06,AHMP08]. A sufficient condition for a kernel to allow for good generalization on a given learning problem is that it induce a large margin of separation between positive and negative classes in its implicit space. In recent work [BBS08,BBS07,BB06] we have developed a theory that more broadly holds for general similarity functions that are not necessarily legal kernel functions. In particular, we give sufficient conditions for a similarity function to be useful for learning that (a) are fairly natural and intuitive (do not require an implicit space and allow for functions that are not positive semi-definite) and (b) strictly generalize the notion of a large-margin kernel function in that any such kernel also satisfies these conditions, though not necessarily vice-versa. We also have partial progress on extending the theory of learning with *multiple* kernel functions to these more general conditions. In this talk we describe the main definitions and results of [BBS08], give our results on learning with multiple similarity functions, and present several open questions about learning good general similarity functions from data.

Multi-Task Learning via Matrix Regularization

Andreas Argyriou, University College London

We present a method for learning representations shared across multiple tasks. The method consists in learning a low-dimensional subspace on which task regression vectors lie. Our formulation is a convex optimization problem, which we solve with an alternating minimization algorithm. This algorithm can be shown to always converge to an optimal solution. Our method can also be viewed as learning a linear kernel shared across the tasks and hence as an instance of kernel learning in which there are infinite kernels available. Moreover, the method can easily be extended in order to learn multiple tasks using nonlinear kernels. To justify this, we present general results characterizing representer theorems for matrix learning problems like the one above, as well as standard representer theorems. Finally, we briefly describe how our method connects to approaches exploiting sparsity such as group Lasso.

Feature Selection: From Correlation to Causality

Isabelle Guyon, Clopinet, Berkeley

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of

internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. This tutorial will cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods. Most feature selection methods do not attempt to uncover causal relationships between feature and target and focus instead on making best predictions. We will examine situations in which the knowledge of causal relationships benefits feature selection. Such benefits may include: explaining relevance in terms of causal mechanisms, distinguishing between actual features and experimental artifacts, predicting the consequences of actions performed by external agents, and making predictions in non-stationary environments.

Learning Bounds for Support Vector Machines with Learned Kernels
Nathan Srebro et al., Toyota Technological Institute & University of Waterloo
Consider the problem of learning a kernel for use in SVM classification. We bound the estimation error of a large margin classifier when the kernel, relative to which this margin is defined, is chosen from a family of kernels based on the training sample. For a kernel family with pseudodimension d_ϕ , we present a bound of $\sqrt{\tilde{O}(d_\phi + 1/\gamma^2)/n}$ on the estimation error for SVMs with margin γ. This is the first bound in which the relation between the margin term and the family-of-kernels term is additive rather then multiplicative. The pseudodimension of families of linear combinations of base kernels is the number of base kernels. Unlike in previous (multiplicative) bounds, there is no non-negativity requirement on the coefficients of the linear combinations. We also give simple bounds on the pseudodimension for families of Gaussian kernels.

Non-sparse Multiple Kernel Learning
Marius Kloft et al., TU Berlin & Fraunhofer Institute FIRST
Approaches to multiple kernel learning (MKL) employ ℓ_1-norm constraints on the mixing coefficients to promote sparse kernel combinations. When features encode orthogonal characterizations of a problem, sparseness may lead to discarding useful information and may thus result in poor generalization performance. We study non-sparse multiple kernel learning by imposing an ℓ_2-norm constraint on the mixing coefficients. Empirically, ℓ_2-MKL proves robust against noisy and redundant feature sets and significantly improves the promoter detection rate compared to ℓ_1-norm and canonical MKL on large scales.

Infinite Kernel Learning
Peter Gehler, Max Planck Institute
In this paper we build upon the Multiple Kernel Learning (MKL) framework. We rewrite the problem in the standard MKL formulation which leads to a Semi-Infinite Program. We devise a new algorithm to solve it (Infinite Kernel Learning, IKL). The IKL algorithm is applicable to both the finite and infinite case and we find it to be faster and more stable than SimpleMKL. Furthermore we present the first large scale comparison of SVMs to MKL on a variety of benchmark datasets, also comparing IKL. The results show two things: a) for many datasets there is no benefit in using MKL/IKL instead of the SVM classifier, thus the flexibility of using more than one kernel seems to be of no use, b) on some datasets IKL yields massive increases in accuracy over SVM/MKL due to the possibility of using a largely increased kernel set. For those cases parameter selection through Cross-Validation or MKL is not applicable.

Kernel Learning for Novelty Detection
John Shawe-Taylor, University College London
We consider kernel learning for one-class Support Vector Machines. We consider a mix of 2- and 1-norms of the individual weight vector norms allowing control of the sparsity of the resulting kernel combination. The resulting optimisation can be solved efficiently using a coordinate gradient method. We consider an application to automatically detecting the appropriate metric for a guided image search task.

POSTER SESSION
-**Ravi S. Ganti, Nikolaos Vasiloglou and Alexander Gray**: Hyperkernel Based Density Estimation

-**Andrew G. Howard and Tony Jebara**: Learning Large Margin Mappings

-**S. Mosci, M. Santoro, A. Verri, S. Villa and L. Rosasco**: A New Algorithm to Learn an Optimal Kernel Based on Fenchel Duality

-**Hua Ouyang and Alexander Gray**: Learning Nearest-Neighbor Classifiers with Hyperkernels

-**Nikolaos Vasiloglou, Alexander G. Gray and David V. Anderson**: Learning Isometric Separation Maps

All other submitted talks are also encouraged to give posters.