Dimensionality Reduction with Contaminated Data: The Very High Dimensional Case
Shie Mannor, Technion - Israel Institute of Technology

We consider the dimensionality-reduction problem (finding a subspace approximation of observed data) for contaminated data in the high dimensional regime, where the number of observations is of the same magnitude as the number of variables of each observation, and the data set contains some (arbitrarily) corrupted observations. We propose a high-dimensional robust principal component analysis (HR-PCA) algorithm that is tractable, robust to contaminated points, and easily kernelizable. The resulting subspace has a bounded deviation from the desired one, and unlike ordinary PCA algorithms, achieves optimality in the case where the proportion of corrupted points goes to zero.

Behind the SNPs: Systematic Synergy-based Analysis of Whole-Genome Associations Data Revisits Apparent Associations
Hani Neuvirth-Telem, IBM Haifa Research Lab

Unraveling the genetic basis of common human diseases is a notable goal expected to have immediate medical applications. A recent leading approach utilizes whole-genome association studies (WGAS). In these studies, common variations for patients and a control group are genotyped via large-scale methodologies. The obtained data are then analyzed using various statistical techniques to pinpoint single nucleotide polymorphisms (SNPs) with a statistically significant association to the disease under investigation. In this work, we utilize the information-theoretic concept of synergy to identify SNP pairs showing a statistically significant association with a trait, which might be overlooked while each SNP is analyzed independently. A detailed examination of the most synergistic SNP pairs reveals a recurring phenomenon, termed here a “context-dependent genotype flip” (CDGF). Specifically, for some SNPs, we observe a systematic genotype bias under specific context of a pairing SNP. Further analysis suggests that synergistic SNP pairs with CDGFs represent a context-dependent genotype error, not observed thus far. Importantly, we argue that such an error may result with various spurious statistical signals in whole genome association studies, yielding misleading conclusions regarding genotype-trait associations.

Testing Properties of Distributions
Ronitt Rubinfeld, Tel-Aviv University

We survey several works regarding the complexity of testing properties of distributions when given access to only a few samples from the distributions. Such properties might include testing if two distributions have small statistical distance; testing various independence properties; testing whether a distribution has a specific shape; and approximating the entropy. Classical techniques, such as the Chi-squared test or the straightforward use of Chernoff bounds, have sample complexities that...
are at least linear in the size of the support of the underlying discrete probability distributions. We will describe bounds for many such testing problems whose sample complexities are sublinear in the size of the support.

**Keynote: Semi-Supervised Online Learning - Train Only When You Need**

*Nicolò Cesa-Bianchi, Università degli Studi di Milano*

An important feature of automatic categorization systems is the ability of computing a level of confidence about the classification of the current instance. A small confidence is viewed as an indication that further training is needed for keeping the desired level of accuracy. In this talk, we describe classification algorithms that occasionally ask for human supervision, achieving a good performance with far less labels than their fully supervised counterparts. Theoretically, these algorithms are shown to dynamically trade off accuracy with number of requested labels, implicitly learning the unknown structure of the data source. We study this trade-off under different assumptions of the data-generation mechanism (from purely stochastic to purely adversarial) and using a simple family of kernel-based classifiers.

**Authorship Attribution in the Wild**

*Moshe Koppel, Bar-Ilan University*

In the vanilla version of the authorship attribution problem, we are asked to assign a long anonymous text to one of a small closed set of candidate authors. This is a straightforward text categorization problem and its solution is simple and well-understood. In the real world (especially in forensic settings), however, we are often faced with attribution problems in which the candidate set might be very large (thousands of candidates) and open (the real author might not be in the candidate set) and in which the training texts and anonymous text might be of limited length. We present a new method, based on randomized feature sets, that solves the real world problem with high precision.

**Semantic Categorization into Very Large Taxonomies: Empirical Investigation**

*Ron Karidi, Israel Innovation Labs, Microsoft*

Given a taxonomy T of semantic concepts, we are interested in semantic classification of new documents into the taxonomy. For a document d and every concept C, we look for p(d,C) which quantifies the semantic proximity of d to C. For flat ontology, previous research considered several IR-distance functions. Our objective is to consider very large T with (non-flat) hierarchical structure. We focused our investigation in the taxonomy offered by ODP, the Open Directory Project (also known as dmoz). The taxonomy consists of a tree structure of ~750,000 categories and sub-categories, with more than four million documents (web pages) populating the nodes. Our work focused on understanding ways to take advantage of the structure when learning the semantic proximity function as well as selection of representatives. This is joint work with Oded Elyada and Liat Segal, both from the Israel Innovation Labs.
Efficient Scalable Algorithms for Structured Prediction
Amir Globerson, Hebrew University

One of the key goals of machine learning is to build classifiers from labeled data. Earlier work in the field focused on cases where labels were binary or confined to a small set of possible classes. However, in recent years, there has been considerable interest in tasks where labels have a more complex structure. For example, labels may correspond to sequences of tags or even combinatorial objects such as trees or matchings. The extension of classical algorithms (e.g., support vector machines) to this scenario presents challenging algorithmic and theoretical challenges.

In this talk, I will describe a general algorithm (called exponentiated gradient) for this setting. The algorithm has an “online” structure (i.e., it processes one sample at a time), but also improves the dual objective at every iteration. I will show how this scheme can be applied to maximum-margin Markov networks and to conditional random fields and analyze its rate of convergence in these cases. I will conclude by showing an application of the algorithm to a large scale natural language parsing task, and demonstrate that it converges significantly faster than the previous state-of-the-art algorithms.

Secrets, Lies, and Email – Information Classification and Identification for Data Loss Prevention (DLP)
Lidor Troyansky, Research Fellow, Websense Inc.

Modern data loss prevention systems are required to classify and identify masses of information in real-time to prevent unauthorized dissemination of sensitive and confidential information. The task is especially challenging since the cost of false positives and false negatives is high, while the definition of “confidential” is, in many cases, subjective. In this talk, I’ll briefly describe the data loss prevention problem, the needs for information classification and identification, and the usage of data-structures known as “Bloom filters” for efficient identification of confidential information.

http://www.haifa.il.ibm.com/Workshops/