Machine Learning Seminar 2007

27 May, 2007

The Knowledgeable Computer: Enhancing Learning Algorithms with Wikipedia-Based Feature Generation

Shaul Markovitch, Technion

When humans approach the task of text categorization, they interpret the specific wording of the document in the much larger context of their background knowledge and experience. On the other hand, state-of-the-art information retrieval systems are quite brittle—they traditionally represent documents as bags of words and are restricted to learning from individual word occurrences in the (necessarily limited) training set. For instance, given the sentence "Wal-Mart supply chain goes real time", how can a text categorization system know that Wal-Mart manages its stock with RFID technology? And having read that "Ciprofloxacin belongs to the quinolones group", how can a machine know that the drug mentioned is an antibiotic produced by Bayer? We propose to enrich document representation through automatic use of a vast compendium of human knowledge---an encyclopedia. We apply machine learning techniques to Wikipedia, the largest encyclopedia to date, which surpasses in scope many conventional encyclopedias and provides a cornucopia of world knowledge. Each Wikipedia article represents a concept and documents to be categorized are represented in the rich feature space of words and relevant Wikipedia concepts. Empirical results confirm that this knowledge-intensive representation brings text categorization to a qualitatively new level of performance across a diverse collection of datasets. This work is done in collaboration with Evgeniy Gabrilovich.

Workstation Capacity Tuning using Reinforcement Learning

Ran Gilad-Bachrach, Intel

Computer grids are complex, heterogeneous, and dynamic systems, whose behavior is governed by hundreds of manually-tuned parameters. As the complexity of these systems grows, automating the procedure of parameter tuning becomes indispensable. In this paper, we consider the problem of auto-tuning server capacity, i.e., the number of jobs a server runs in parallel. We present three different reinforcement learning algorithms, which generate a dynamic policy by changing the number of concurrent running jobs according to the job types and machine state. The algorithms outperform manually-tuned policies for the entire range of checked workloads, with average throughput improvement greater than 20%. On multi-core servers, the average throughput improvement is approximately 40%, which hints at the enormous improvement potential of such a tuning mechanism with the gradual transition to multi-core machines.

This is a joint work with Aharon Bar-Hillel, Amir Di-Nur, Liat Ein-Dor and Yossi Ittach.

IBM Haifa Leadership Seminar – Abstracts

Continuous Time Markov Networks

Nir Friedman, Hebrew University, Jerusalem

A central task in many applications is reasoning about processes that change over continuous time. We can formally reason about such processes as Continuous Time Markov Processes. However, in most applications, we deal with structured state space that involves exponential number of states. This raises the challenge of representation, inference and learning with such structured continuous time processes. In this talk, I will describe continuous time Markov networks (CTMNs), a representation language for continuous-time dynamics, particularly appropriate for modeling biological and chemical systems. In this language, the dynamics of the process is described as an interplay between two forces: the tendency of each entity to change its state, which we model using a continuous-time proposal process that suggests possible local changes to the state of the system at different rates; and a global fitness or energy function of the entire system, governing the probability that a proposed change is accepted, which we capture by a Markov network that encodes the fitness of different states. We show that the fitness distribution is also the stationary distribution of the Markov process, so that this representation provides a characterization of a temporal process whose stationary distribution has a compact graphical representation. I will describe the semantics of the representation, its basic properties, and algorithms for learning and inference with such models.

This is joint work with Tal El-Hay, Raz Kupferman, and Daphne Koller.

Feature Selection for Categorical Features with Many Values

Sivan Sabato, IBM Haifa Research Lab

Feature selection is an important task in machine learning with various applications. One approach for feature selection is ranking features based on a criterion which, intuitively, assesses the relevance of each feature to the label. Several relevance criteria have been previously proposed in the literature, including, for instance, the Gini index and the misclassification error. While these criteria have been found to be effective in many cases, they are likely to fail in the presence of categorical features with a large number of possible values. We propose a different approach to this problem and provide a new ranking criterion, for which we prove accuracy bounds. These bounds are independent of the number of possible values a feature may take or their distribution. As a result, the new criterion can be used in feature selection to avoid the problems faced by other ranking criteria.

This is joint work with Shai Shalev Shwartz.

Maximum Entropy and Species Distribution Modeling

Robert Schapire, Princeton University

Modeling the geographic distribution of a plant or animal species is a critical problem in conservation biology. To save a threatened species, one first needs to know where it prefers to live and what its requirements are for survival. From a machine-learning perspective, this is an especially challenging problem in which the learner is presented with no negative examples and often only a tiny number of positive examples. In this talk, I will describe the application of maximum-entropy methods

to this problem, a set of decades-old techniques that happen to fit the problem very cleanly and effectively. I will describe a version of maxent that we have shown enjoys strong theoretical performance guarantees, which enable it to perform effectively even with a very large number of features. I will also describe some extensive experimental tests of the method, as well as some surprising applications.

This talk includes joint work with Miroslav Dudik and Steven Phillips.

Global Learning with Constraints

Dan Roth, University of Illinois at Urbana-Champaign

The maturity of machine learning techniques allows us today to learn many low level predicates and to generate an appropriate vocabulary over which reasoning methods can be used to make significant progress in higher level domain decisions. I will describe research on a framework that combines learning and inference, and exhibit its use in the natural language processing domain. Key in this framework is the ability to incorporate declarative and expressive global information into the learning and decision stage. I will discuss the use of this framework as (1) a way to allow the output of local classifiers for different problem components to be assembled into a whole that reflects global preferences and constraints; (2) a way to improve probabilistic models by enforcing additional expressive constraints and (3) a way to significantly improve semi-supervised training of structured models. Examples will be drawn from 'wh' attribution in natural language processing (determining who did what to whom when and where) and from information extraction problems.

Support Vector Machine Solvers: Large-Scale, Accurate, and Fast (Pick Any Two)

Elad Yom-Tov, IBM Haifa Research Lab

Support vector machines (SVMs) have proved to be highly successful for use in many applications that require classification of data. However, training an SVM requires solving an optimization problem that is quadratic in the number of training examples. This is increasingly becoming a bottleneck for SVMs because, while the size of the datasets is increasing especially in applications such as bioinformatics, single-node processing power has leveled off in recent years. One possible solution to these trends lies in solving SVMs on multiple computing cores or on computing clusters. In my talk I will show two approaches to solving SVMs in parallel, one based on holding the complete kernel matrix in distributed memory and the other on training an ensemble of smaller SVMs in parallel. I will compare these solvers to a popular single-node solver. The comparison covers accuracy, speed, and the ability to process large datasets. I will show that while none of these solvers performs well on all three metrics, each of them ranks high on two of them. Finally, I will describe IBM's Parallel Machine Learning toolbox, which allows practitioners to rapidly implement parallel learning algorithms.

Project URL: **http://www.alphaworks.ibm.com/tech/pml**

The Uncertainty Principle of Cross-Validation

Mark Last, Ben-Gurion University of the Negev, Beer-Sheva

In machine learning and data mining, we often have to deal with data sets of limited size due to economic, timing and other constraints. Usually our task is two-fold: to induce the most accurate model from a given dataset and to estimate the model's accuracy on future (unseen) examples. Cross-validation is the most common approach to estimating the true accuracy of a given model and it is based on splitting the available sample between a training set and a validation set. Practical experience shows that any cross-validation method suffers from either an optimistic or a pessimistic bias in some domains. In this talk, we present a series of large-scale experiments on artificial and real-world datasets, where we study the relationship between the model's true accuracy and its cross-validation estimator. Two stable classification algorithms (ID3 and info-fuzzy network) are used for inducing each model. The results of our experiments have a striking resemblance to the well-known Heisenberg Uncertainty Principle: the more accurate is a model induced from a small amount of real-world data, the less reliable are the values of simultaneously measured cross-validation estimates. We suggest calling this phenomenon "the uncertainty principle of cross-validation".