Statement of Research

Marina Meilă

Objective

My main interest is learning models from data, in a statistical framework. I am using machine learning, belief networks (or graphical probability models) and algorithmic graph theory to design models and algorithms for automated learning, inference and decision making that are computationally efficient and thus applicable to large real-world problems. I am interested in domain specific learning paradigms for information retrieval, language, vision and data mining tasks.

Most of my projected research is, in one way or another, related to the field of graphical models. A graphical model describes a multidimensional domain by focusing on the relevance/irrelevance relationships between the variables (in probabilistic terms they are dependencies/independencies). These relationships are represented as a graph called the *structure* of the model. A graphical model thus separates the dependency structure from its detailed functional form; this makes for an excellent support for human intuition and enables one to construct general inference and learning algorithms for this model class.

Directions for research

Mixtures of trees. One of the simplest graphical model is the tree. Although the tree distribution is not novel, it is only recently that it has started to receive attention in machine learning. My paper "Estimating dependency structure as a hidden variable" was one of the earliest contributions to this. Additionally, it showed how to combine tree models into more powerful models (i.e. mixtures), establishes the Bayesian framework for learning trees and mixtures and demonstrates their properties as feature selectors in classification. The tree is singular among belief nets because its structure can be learned efficiently; thanks to this property the mixture of trees overcomes one of the main difficulties in applying Bayes nets techniques to large problems: the fact that learning the structure of a Bayes net from data is an intractable task. From here on research can be pursued in several directions:

Learning the structure of a belief network by variable partitioning. The idea is to attack the structure learning problem top-down, in a divide and conquer approach, whereby the space of all variables is partitioned into conditionally independent subsets on which the structure search can be performed separately. This type of technique, applied recursively, can be seen as a divide-and-conquer approach to structure learning. The advantage w.r.t. established structure learning methods is computational: the latter methods need to explicitly evaluate each structure individually whereas a top-down method can quickly eliminate large portions of the search space making itself potentially useful for higher dimensional problems.

Fast algorithms. I have recently become interested in learning methods that scale well with the dimension of the domain. One result is that I developed a method for learning trees over discrete domains that can be very fast when the data are, in some sense, sparse. The algorithm is jointly subquadratic in the number of variables and the size of the data set (while the standard version of this learning algorithm is jointly cubic). The speedup, which can attain 3 orders of magnitude makes trees and mixtures thereof a tractable option for modeling higher dimensional domains. I intend to apply this algorithm to information retrieval and document classification, and to other data mining tasks. I am now working on randomized algorithms for learning trees and Markov Random fields for modeling geophysical data. The motivation for this line of work is the fact that I see trees and tractable belief networks as building blocks for modeling domains with large number of variables and unknown dependency structure. These models are not necessarily attempting to find the "true" structure of the domain, but are able to provide good approximate models for automated inference tasks. Further reaching applications of these results are in active learning and pattern recognition.

Incremental density estimation. The work on mixtures can be seen as a first step towards the broader objective of constructing complex density models from simple ones. The idea of my approach is that, instead of fitting a model of fixed complexity (e.g. a mixture with a given number of components), to fit a simple model first, then improve it by learning another simple model that corrects the errors of the first one, and so on, incrementally. In other words, to do for density estimation what boosting does for classification, additive models do for regression or the Fourier decomposition does for signals.

Background

My previous interests encompass training associative memories (an original data dependent criterion for the capacity of an associative memory), reinforcement learning, dynamic programming and Markov decision processes, transfer of learning and combinations of mixtures of experts and hidden Markov models. Besides the areas of direct interest, I have followed the major developments in the field of machine learning.

In the field of graphical models I worked first on algorithms for optimal triangulation (a new criterion function and an original method for belief net triangulation), then I turned to learning algorithms: I derived the mixture of trees learning algorithm and its extensions. Among these extensions: the method for learning structures with hidden variables via mixtures of trees.

I received a strong mathematical preparation at the Bucharest Polytechnic University, where I obtained my MS degree, that I supplemented by my own readings. During that time I twice took part and won first place in the Institute's (21,000 students) mathematics competition; I also received a national award in the same competition. My engineering studies focused on linear system theory, signal processing, numerical methods and optimization. At MIT, my minor is in mathematics. I have become acquainted with parts of graph theory due to my research in graphical models.