Learning with Mixtures of Trees

by

Marina Meilă-Predoviciu

M.S. Automatic Control and Computer Science (1985)
Universitatea Politehnica Bucureşti

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1999

Signature of Author . . . .
Department of Electrical Engineering and Computer Science
January 14, 1999

Certified by . . . .
Michael I. Jordan
Professor, Department of Brain Sciences
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Learning with Mixtures of Trees
by
Marina Meilă-Predoviciu

Revised version of a thesis submitted to the
Department of Electrical Engineering and Computer Science
on January 14, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

One of the challenges of density estimation as it is used in machine learning is that usually the data are multivariate and often the dimensionality is large. Operating with joint distributions over multidimensional domains raises specific problems that are not encountered in the univariate case. Graphical models are representations of joint densities that are specifically tailored to address these problems. They take advantage of the (conditional) independencies between subsets of variables in the domain which they represent by means of a graph. When the graph is sparse, graphical models provide an excellent support for human intuition and allow for efficient inference algorithms. However, learning the underlying dependence graph from data is generally NP-hard.

The purpose of this thesis is to propose and to study a class of models that admits tractable inference and learning algorithms yet is rich enough for practical applications. This class is the class of mixtures of trees models. Mixtures of trees inherit the excellent computational properties of tree distributions (themselves a subset of graphical models) but combine several trees in order to augment their modeling power, thereby going beyond the standard graphical model framework.

The thesis demonstrates the performance of the mixture of trees in density estimation and classification tasks. In the same time it deepens the understanding of the properties of the tree distribution as a multivariate density model. Among others, it shows that the tree classifier implements an implicit variable selection mechanism.

Learning mixtures of trees from data is a central subject of this thesis. The learning algorithm that is introduced here is based on the the EM and the Minimum Weight Spanning Tree algorithms and is quadratic in the dimension of the domain.

This algorithm can serve as a tool for discovering hidden variables in a special but important class of models where, conditioned on the hidden variable, the dependencies between the observed variables become sparse.

Finally, it is shown that in the case of sparse discrete data, the original learning algorithm can be transformed in an algorithm that is jointly subquadratic and that in simulations achieves speedups factors of up to a thousand.

Thesis Supervisor: Michael I. Jordan
Title: Professor, Department of Brain Sciences

Acknowledgments

The years at MIT have been a fantastic experience and it is a joy to remember those who concurred in making it so and to thank those whose influence contributed to this thesis.

Michael Jordan, my thesis advisor, fostered my interest in reinforcement learning, statistics, graphical models, differential geometry and . . . who knows what next? He was the soul of the unforgettable "Jordan lab" and of its equally unforgettable group meetings. He supported me materially and morally during all these years while giving me complete freedom in pursuing my work. In spite of this freedom, my thesis owes enormously to his inspiring influence. I have deeply appreciated his relentless enthusiasm for learning and deepening one's understanding, the high standards he set, the genuinely positive attitude enriched by a sense for beauty that is his approach to science and life.

I thank Alan Willsky and Paul Viola, the two other members of my committee, for the interest and support they gave to this project and especially for asking just the right questions to make this thesis a better one. I very much hope that these discussions will be continue. Paul suggested a direction of work which resulted in the accelerated tree learning algorithms. The same algorithms benefited greatly from a series of very stimulating meetings with David Karger. Eric Grimson and Alvin Drake gave me good advice from my fist days on, Tomaso Poggio and his group "adopted" me during my last semester of studentship, Peter Dayan, to whom I should have talked more, found the time to listen to me and an insightful comment on every subject that I could have brought up.

Quaid Morris shared my enthusiasm for trees, coded the first mixtures of trees program, did the digits and random trees experiments. David Heckerman and the Microsoft DTAS group hosted me in Seattle for four intense months, forcing me to take a distance from this work that proved to be entirely to its benefit. David with Max Chickering provided code and ran part of the Alarm experiments.

Carl de Marcken was the most insightful critic of my ideas and my writing, answered my questions and thought me uncountably many things I needed to know, pondered with me over dozens of problems as if they were his own and was always the first to share with me the joy of a success.

I have had countless opportunities to appreciate the quality of the education I received at my Alma Mater, the Polytechnic University of Bucureşti; I take this one to thank Vasile Brînzănescu, Paul Flondor, Dan Iordache, Theodor Dănilă, Corneliu Popeea, Vlad Ionescu, Petre Stoica for everything they thought me.

I chose the thesis topic, but chance chose my officemates. I couldn't have been luckier: Greg Galperin, Tommi Jaakkola, Thomas Hofmann listened to my half baked ideas, answered patiently thousands of questions and let themselves be engaged in exciting discussions on every topic under the sun. They are now among my best friends and one of the reasons I find myself enriched by the years at MIT.

Finally, there is the circle of love that constantly and invisibly surrounds us with its strength. Gail, Butch, Paya and Natasha who welcomed me in their family. Felicia. Carl. My mother – teacher, friend and partner in adventures. My grandfather, Judge Ioan I. Predoviciu, to whose memory I dedicate this thesis.

This research was supported through the Center for Biological and Computational Learning at MIT funded in part by the National Science Foundation under contract No. ASC-92-17041.

Contents

List of Figures

List of Tables

Index of Algorithms