Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

# Learning with Mixtures of Trees

**Marina Meilă** MMP@STAT.WASHINGTON.EDU
*Department of Statistics*
*University of Washington*
*Seattle, WA 98195-4322, USA*

**Michael I. Jordan** JORDAN@CS.BERKELEY.EDU
*Division of Computer Science and Department of Statistics*
*University of California*
*Berkeley, CA 94720-1776, USA*

**Editor:** Leslie Pack Kaelbling

## Abstract

This paper describes the mixtures-of-trees model, a probabilistic model for discrete multidimensional domains. Mixtures-of-trees generalize the probabilistic trees of Chow and Liu (1968) in a different and complementary direction to that of Bayesian networks. We present efficient algorithms for learning mixtures-of-trees models in maximum likelihood and Bayesian frameworks. We also discuss additional efficiencies that can be obtained when data are "sparse," and we present data structures and algorithms that exploit such sparseness. Experimental results demonstrate the performance of the model for both density estimation and classification. We also discuss the sense in which tree-based classifiers perform an implicit form of feature selection, and demonstrate a resulting insensitivity to irrelevant attributes.

**Keywords:** Bayesian Networks, Mixture Models, Chow-Liu Trees

## 1. Introduction

Probabilistic inference has become a core technology in AI, largely due to developments in graph-theoretic methods for the representation and manipulation of complex probability distributions (Pearl, 1988). Whether in their guise as directed graphs (Bayesian networks) or as undirected graphs (Markov random fields), *probabilistic graphical models* have a number of virtues as representations of uncertainty and as inference engines. Graphical models allow a separation between qualitative, structural aspects of uncertain knowledge and the quantitative, parametric aspects of uncertainty...

*Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.*

## Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

## Appendix A.

In this appendix we prove the following theorem from Section 6.2:

**Theorem** *Let $u, v, w$ be discrete variables such that $v, w$ do not co-occur with $u$ (i.e., $u \neq 0 \Rightarrow v = w = 0$ in a given dataset $\mathcal{D}$). Let $N_{v0}, N_{w0}$ be the number of data points for which $v = 0, w = 0$ respectively, and let $I_{uv}, I_{uw}$ be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then*

$$N_{v0} \; > \; N_{w0} \;\; \Rightarrow \;\; I_{uv} \; \leq \; I_{uw}$$

*with equality only if $u$ is identically 0.* ∎

**Proof**. We use the notation:

$$P_v(i) \; = \; \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \; \equiv \; P_v(0) \; = \; 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of $v$ taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by $H$. We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0$....

*Remainder omitted in this sample. See http://www.jmlr.org/papers/ for full paper.*

## References

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman Publishers, San Mateo, CA, 1988.