Dirichlet Process Mixtures of Trees

Sergey Kirshner* and Padhraic Smyth†

*AICML, Department of Computing Science, University of Alberta, Canada
†Donald Bren School of Information and Computer Science, University of California, Irvine, USA

Problem Description

Given: a set \( \{ \mathbf{x} \} \) of i.i.d. vector samples from a multivariate categorical-valued distribution \( p \)

Want: an approximation to \( p \) for both descriptive and generative purposes

Why? medical and biological domains, vision, OCR, multi-site precipitation modeling

How? Bayesian networks, MRFs (both very slow if the structure is not known), mixtures of trees

Tree-dependent Distribution

Finite mixture with a tree-dependent distribution for each component:

\[ q(\mathbf{x}) - \sum_{\mathbf{\theta}} p^T(\mathbf{x}) \]

Efficient \( (O(Nd^2b^2)) \)

Restrictive and inaccurate

Use Bayesian approach to average over the tree structures, parameters, and the number of mixture components.

Conjugate Prior for Tree-dependent Distributions

Spanning Trees

Assign a non-negative weight \( \beta_{uv} \) for each potential edge \( \{u,v\} \). Define a probability distribution over spanning trees with these weights as hyperparameters (denoted by \( \beta \)).

Efficient \( (O(NKd^2b^2)) \) per EM iteration

Still restrictive, \( K = ? \)

Use Bayesian approach to average over the tree structures, parameters, and the number of mixture components.

Collapsed Gibbs Sampling

Estimating parameters from a synthetic data set simulated from a \( 6 \)-state mixture of trees. DPMT outperforms the estimated model with the true functional form.

Parameters

Assume given a pseudo-set over the same domain as \( \mathbf{x} \). Univariate and bivariate pseudo-counts serve as hyperparameters (denoted by \( \Psi \) to the structured Dirichlet prior:

Efficient \( (O(db)) \) Time complexity of computation of partition function \( Z \): \( O(d^3) \).

Time complexity for sampling of \( \theta \): \( O(d^3) \) randomized or \( O(d^3) \) otherwise.

Evaluation

Estimating the number of weather states (mixture components) for Ceará region data.

DPMT outperforms finite mixtures of trees for these data sets.