Bayesian Unsupervised Learning for High-dimensional Distributions with Tree-based Methods

Naoki Awaya
School of Political Science and Economics, Waseda University

Abstract

Estimating distributional structures, such as density estimation and two-sample comparison, is a fundamental task in data science. However, estimating high-dimensional distributions is widely recognized as challenging due to the well-known curse of dimensionality. In the case of supervised learning, where one needs to estimate an unknown function often defined on a high-dimensional space, a common Bayesian approach is to use tree-based methods such as the Classification and Regression Tree (CART) and the Bayesian Additive Regression Tree (BART). These methods are known to be effective for such challenging tasks with feasible computation costs. This presentation aims to introduce their counterparts for unsupervised learning.

First, we discuss the generalization of a Bayesian non-parametric model called the Polya-tree process, which is a tree-based model for estimating distributions. Its usage was previously limited to exploring relatively low-dimensional structures due to the lack of scalable posterior computation methods, especially for the posterior of tree structures. To address this problem, we propose a new sequential Monte Carlo algorithm to approximate the posterior with substantially smaller computation costs by efficiently exploring the tree space. Introducing this efficient sampling algorithm also enables the use of a more flexible Bayesian model with weaker restrictions on trees, which improves predictive performance. The proposed method is demonstrated on high-dimensional biological data consisting of two groups, and we show that the differences between the groups is successfully captured.

Next, we discuss a generalization of the proposed method by defining an ensemble of trees. As widely recognized in the context of supervised learning, constructing an ensemble, which is a combination of multiple tree structures, substantially improves numerical performance. Motivated by this, we propose an ensemble of Polya tree-based models. To this end, we introduce a transformation named "tree-CDF," which generalizes cumulative distribution functions (CDFs) and defines a combination of multiple distributions in terms of a composition of their corresponding tree-CDFs. This mathematical structure can be seen as a version of machine learning network models called normalizing flows. We show that the ensemble based on tree-CDFs enables the introduction of an efficient algorithm, considered an unsupervised counterpart of the boosting algorithm for regression and classification. Numerical studies demonstrate the efficiency of the proposed boosting algorithm over existing normalizing flow methods.