THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Statistics Colloquium

Joint colloquium with the Data Science Institute (DSI)

Li Ma

Department of Statistical Science, Department of Biostatistics & Bioinformatics
Duke University

"Two examples of scalable unsupervised learning with trees and recursive partitions"

Monday, January 29, 2024, at 11:30 AM

Jones 303, 5747 S. Ellis Avenue

Refreshments will be served prior to the Seminar at 11:00 am in Jones 303

Abstract

Trees and recursive partitions are most well-known in supervised learning for predictive tasks, such as regression and classification. Famous examples include CART and its various forms of ensembles—e.g., random forest and boosting. A natural question is whether such successes can be replicated in the context of unsupervised problems. I present two recent examples of tree-based approaches in the context of unsupervised learning, where the primary objective is to learn the underlying nature of complex multivariate, possibly high-dimensional distributions based on unlabeled i.i.d. training data. In both examples, the employment of trees and partitions leads to highly efficient, statistically rigorous algorithms that scale approximately linearly in the sample size. These algorithms can be trained quickly or even in real-time on single computers. The first example addresses density estimation and generative modeling in multivariate sample spaces using an additive ensemble of tree-based density learners. The method is a counterpart of supervised tree boosting and preserves many desirable properties of its supervised cousin. It also has a close connection to the so-called normalizing flows based on sequentially transforming a base distribution to obtain a desired distribution and so produces a generative model that can be directly sampled from through sequential inverse transforms. The second example involves testing and learning dependency structures between random variables (or vectors). In this context utilizing trees and partitions leads to a generalization of the classical Fisher exact test on 2 by 2 tables to general multivariate sample spaces, while maintaining the ability to ensure exact finite-sample validity, thereby avoiding the need for either asymptotic approximation or resampling. It also provides a means to identifying the nature of the underlying dependency as opposed to merely testing the hypothesis of independence, which has been the focus of most existing methods but rarely holds exactly in high-dimensional applications.

Information about building access for persons with disabilities may be obtained in advance by calling Shannon Kokesh, Department Secretary, at 773-702-8333. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.