How to Simulate Conditionally Specified Models: Solutions of a Computational Problem in Statistics

Yuchung J Wang
Rutgers University–Camden, New Jersey USA

Abstract

Traditional machine learning (ML) chooses from a collection of algorithms to solve a specific problem. Algorithms that minimizes the prediction error include decision trees, random forest, boosting, K-means clustering, SVM, HMM, Kalman filter, linear regression, Boltzmann mechanic etc. A conceptually different approach is the model-based ML, which uses probabilistic graphical models (GM) for inference and prediction. There are two kinds of GM: Bayes network and dependence network, where the former is represented by a directed acyclic graph (DAG) and the latter by a directed cyclic graph (DCG).

Both DAG and DCG use conditioning to propagate probabilities throughout the network, but with a distinctive difference. In a DAG, all of the propagations move in one direction, from children nodes to parent nodes, whereas DCG allows the propagations to have a feedback circle. Such conditionally specified DCG models offer several advantages over the more conventional joint models and DAG. In recent literature of multiple imputation, the conditional approach has become more popular than the joint approach. Basically, we use the complete data to propose a model for every variable with missing data via regression or logistic regression. Then, use the Gibbs sampling to generate imputations for the missing data. However, it would be restrictive to require that every regression/classification must involve the same set of variables. Feature selection often reduces the set of predictors, thus make the regressions local. A mixture of full and local conditionals is referred to as a *partially collapsed Gibbs sampler (PCGS)*, which was invented to achieve faster convergence via reduced conditioning. However, we show that its implementation requires choosing a correct scan order. Using an invalid scan order will bring about an incorrect transition kernel, which leads to the wrong stationary distribution. We prove a necessary and sufficient condition for PCGS to correctly sample the joint distribution. We propose an algorithm that identifies all of the valid scan orders for a given DCG model. We will use several examples to illustrate the faster convergence of a PCGS, and the difficulty of verifying convergence of MCMC visually. In addition, our method offers a new compatibility check for conditionals of different localities.

Key words: computational problem in statistics; dependence network; faster convergence; Markov chain Monte Carlo; Model-based machine learning; partially collapsed Gibbs sampler.