A Reference distribution for the Kullback-Leibler distance for Bayesian network conditional probability tables.

Russell G. Almond and Seyfullah Tingir
Florida State University

December 15, 2016

Abstract

A discrete Bayesian network is defined by a collection of conditional probability tables indexed by the nodes in an acyclic directed graph. For each node (variable) X the model specifies the conditional probability of X give $pa(X)$, the parents of X in the graph. Thus, to estimate a Bayesian network, a statistician must estimate a collection of conditional probability distributions $\tilde{p}(X|pa(X))$ for each node in the graph. A natural procedure for evaluating a parameter estimation procedure for a Bayesian network is to pose a collection of distributions $p(X|pa(X))$, generate random data from that collection, and estimate the the distributions $\tilde{p}(X|pa(X))$ from that data. The difference between the original, $p(X|pa(X))$, and estimated, $\tilde{p}(X|pa(X))$, distributions is then a measure of the quality of the estimation algorithm.

A frequently used measure of distance between two distributions is the Kullback-Leibler distance:

$$D_{KL}\left(p(X|pa(X)) \parallel \tilde{p}(X|pa(X))\right) = \sum_{j \in pa(X)} \sum_{x \in X} \log \frac{p(X|pa(X))}{\tilde{p}(X|pa(X))} p(X|pa(X)). \tag{1}$$

Technically speaking, each row of the table is a probability distribution, so the measure in Equation 1 is a sum of Kullback-Leibler divergences. Although this measure is commonly used, the expected size depends on the application. Our paper describes a simple simulation experiment showing the expected size of the Kullback-Leibler divergence for a number of common conditional probability table sizes.

Key words: Bayesian Networks, Kullback-Leibler Distribution