Score-Based Learning Algorithms

Question

What types of algorithms does BayesiaLab use for learning network structures?

Answer

We have developed our proprietary s*core-based* learning algorithms. As opposed to the *co nstraint-based* algorithms that use independence tests to add or remove arcs between nodes, we utilize the *MDL* score (*Minimum Description Length*) to measure the quality of candidate networks with respect to the available data.

This score, which is derived from Information Theory, allows to automatically take into account the data likelihood with respect to the network and the structural complexity of the network.

Minimum Description Length

The *Minimum Description Length (MDL)* score is a two-component score that has been traditionally used in the Artificial Intelligence community for estimating the number of bits required to represent a model and the data given this model. For structural learning of Bayesian networks, the model is the Bayesian network (graph plus probability tables), whereas the number of bits for representing the data given the Bayesian network is inversely proportional to the probability of the observations returned by the model.

$$MDL(B,D) = \alpha DL(B) + DL(D|B)$$

where:

- α represents the BayesiaLab Structural Coefficient (the default value is 1), a parameter that allows changing the weight of the MDL structural part (the lower its value, the greater the complexity of the resulting networks),
- DL(B) the number of bits to represent the Bayesian network *B* (graph and probabilities), and
- *DL(D|B)* the number of bits to represent the dataset *D* given the Bayesian network *B*.

Bayesian Network Part

DL(B)

The number of bits to represent a Bayesian network is equal to the number of bits to represent the structure plus the number of bits to represent the probability distributions.

$$DL(B) = DL(G) + DL(P|G)$$

where *G* refers to the Graphical structure, and *P* to the set of Probability tables.

DL(G)

The coding of the structure implies the identification of each node plus its parents.

$$DL(G) = \sum_{i}^{n} \left(log_2(n) + log_2 \binom{n}{||\pi_i||} \right)$$

where

- *n* is the number of random variables (nodes): $X_1, ..., X_n$
- π_i is the set of the random variables that are parents of X_i in the graph *G*
- and $||\pi_i||$ is the number of parents of random variable X_i.

DL(P|G)

The number of bits to represent the probability distributions is proportional to the number of cells of the conditional probability tables.

$$DL(P|G) = \sum_{i}^{n} \left(\prod_{j}^{||\pi_i||} S_j \times (S_i - 1) \times DL(p) \right)$$

where

- S_i is the number of states of random variable X_i,
- *p* is the probability associated with the cell.

As this probability is not known prior to learning the network, we are using the following classical heuristic:

$$DL(p) = \frac{log_2(N)}{2}$$

where N is the number of observations in the data set.

Data Part

The number of bits for representing the data given the Bayesian network is inversely proportional to the probability of the observations returned by the model.

$$DL(D|B) = \sum_{j=1}^{N} DL(e_j|B)$$

$$DL(D|B) = \sum_{j=1}^{N} log_2(\frac{1}{P_B(e_j)})$$

$$DL(D|B) = -\sum_{j=1}^{N} log_2(P_B(e_j))$$

where

- e_j is the *n*-dimensional observation described in row *j,* and
- $P_B(e_j)$ is the joint probability of this observation returned by the Bayesian network *B*.

The chain rule allows rewriting this equation with:

$$DL(D|B) = -\sum_{j=1}^{N} log_2 \left(\prod_{i=1}^{n} P_B(x_{ij}|\pi_{ij}) \right)$$

$$DL(D|B) = -\sum_{j=1}^{N} \sum_{i=1}^{n} log_2 \left(P_B(x_{ij}|\pi_{ij}) \right)$$

For each candidate Bayesian network, which is generated during the search and evaluated with the *MDL* score, the corresponding parameters are computed using *Maximum Likelihood Estimation.*