情報意味論 (第8回補足)
ベイジアンネットワークの学習

慶應義塾大学理工学部
櫻井 彰人

BNの学習

BNをデータから構成する方法に2種類ある:

- 制約を発見していく方法
  - 統計的検定を行って、条件付独立な変数組を発見していく
  - これを満たす DAG を見つける
- スコア関数を用いる方法
  - DAG を比較するスコア関数を用いる,
    eg. Bayesian, BIC, MDL, MML
  - データに最もよくfitする DAG を選ぶ

注: 通常、Markov等価性(説明してありません)による制約を考える。というのも、Markov等価なDAGは統計的には区別できないからである.

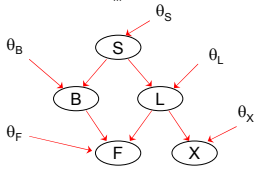
Bayes的方法(1)

(Cooper and Herskovits, 1992)

データを用いて、条件付独立性に関する統計的推定を行う
- 確率的関係をよりよく表現するモデルを探す

M – 構造を表す離散確率変数. 値 m はありうる DAG 構造.
Mの値は分布するとする. 確率分布を P(m) で表す.

Θ_m – モデル m に対応した連続ベクトル値の確率変数(パラメータ). 値 θ_m はそのパラメータ値. Θ_m の値も分布する. 確率分布を $P(\theta_m \mid m)$ で表す.



G.F. Cooper and E. Herskovits (1992)
Machine Learning, 9, 309-47

Bayes的方法(2)

訓練データ集合を D, DAG構造 m の事後確率は, D が与えられたとして:

$$P(m \mid D) = \frac{P(m)P(D \mid m)}{\sum_{m'} P(m')P(D \mid m')}$$

但し

$$P(D \mid m) = \int P(D \mid \theta_m, m)P(\theta_m \mid m)d\theta_m$$

は周辺尤度である. 例によって事前分布 P(m) が一様分布であれば

$$P(m \mid D) \propto P(D \mid m)$$

従って、尤度最大化は事後確率最大化となる.

Bayes的方法 (3)

Cooper and Herskovits (1992) によれば、周辺尤度は次の通り

$$P(D \mid m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

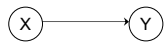
n – 全ノード数
q_i – ノード X_i の親ノード達の値全部の組合せ総数
r_i – ノード(離散確率変数) X_i の値の総数
α – 事前分布である Dirichlet 分布のパラメータ(i はノード, $1 \le j \le q_i$)
N – データ数. ノードi, 親ノード値の組合せj, k番目の値

この P(D | m) は Bayesian scoring function として知られている.

G.F. Cooper and E. Herskovits (1992)
Machine Learning, 9, 309-47

計算例

次の DAG m_1 と訓練データ D を考える



データID	X	Y
1	1	1
2	1	2
3	1	1
4	2	2
5	1	1
6	2	1
7	1	1
8	2	2

$P(D \mid m_1)$ は

$$P(D \mid m_1) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

Y (i=2) に対し $q_i = 2$ (X は2値) かつ $r_i = 2$ (Yは2値). j = 1 に対応する項は

X=1 Y=1 Y=2

$$\frac{\Gamma(2)}{\Gamma(2+5)} \frac{\Gamma(1+4)}{\Gamma(1)} \frac{\Gamma(1+1)}{\Gamma(1)}$$

他の項も計算すれば $P(D \mid m_1) = 7.22 \times 10^{-6}$

R.E. Neapolitan, Learning Bayesian Networks (2004)

計算例 (続)

m_1 は、変数 X と Y の間に(条件付)独立性がないことを示すDAG (の Markov同値クラス)の代表と考えることができる.

m_2 を エッジがない DAG とすると $P(D \mid m_2) = 6.75 \times 10^{-6}$

X　　Y

さらに m_1 と m_2 の事前確率は等しい、すなわち $P(m_1) = P(m_2) = 0.5$ とすると m_1 の事後確率は m_2 の事後確率より大きくなる.

Bayesの定理により

$$P(m_1 \mid D) = \frac{P(D \mid m_1)P(D \mid m_1)}{P(D \mid m_1)P(D \mid m_1) + P(D \mid m_2)P(D \mid m_2)} = \frac{7.215 \times 0.5}{7.215 \times 0.5 + 6.7465 \times 0.5} = 0.517$$

探索アルゴリズムの必要性

理想的には全DAGの空間を網羅的に探索し、前述の Bayesian scoring function を最大化するDAGを見つけたい.

しかし、ノード数を大きく(ほんの少し大きく)しただけで、DAGの数は莫大なものとなる:

ノード数	DAG総数
1	1
2	3
3	25
4	543
5	29,281
10	4.2×10^{18}

様々な発見的方法が開発されている

K2 Algorithm (1)

(Cooper and Herskovits, 1992)

n 変数 $\{X_1, X_2, \ldots, X_n\}$ 間に順序があると仮定する。すなわち, $j > i$ ならば、X_j は X_i の親にはなれないとする.

X_2 について

X_2 に親がないとしてBayesian score を求める

X_2 の親が X_1 として Bayesian score を求める. これがより大きければ X_1 から X_2 へのエッジをつける.

X_iについて

X_i に親がないとして Bayesian score を求める

X_i に親が一つだとして Bayesian scoreを求める. 親がない場合より大きいscoreがあればその最大値を与える X_j からのエッジをつける.

次に第二番目の親を選んで同様のことを試みる. これをscoreが大きくならないまで続ける.

K2 Algorithm (2)

変数の順序を {X, Y, Z} とする