「ベイジアンネットワーク概説」

4.3 確率推論アルゴリズムの計算量
4.4 不完全データ・欠損データからの EM 学習

新納浩幸

推論アルゴリズムの計算時間の比較

LoopyBP、JT (Junction Tree)、SS(Systematic Sampling)を比較

ノード数	LoopyBP	JT	SS
20	119 ms	112 ms	445 ms
50	314 ms	997 ms	1845 ms
100	2.283 sec	10.820 sec	4.197 sec
300	4.765 sec	実行不可	20.367 sec

比較実験の結論

LoopyBP

ネットワークが大規模でも高速、メモリ消費も少ない

解の精度、収束性に問題ある場合もある

JT

ノード数が多いと実行不可能

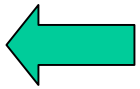
SS

ネットワークが小さいときサンプル数を十分取れば
高い精度の解が得られる
ネットワークが大規模だとメモリ消費量、計算時間が問題

完全データと不完全データ

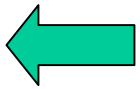
データから CPT を推論するとき

全ての組み合わせについてのデータが存在



完全データ

一部の組み合わせについてのデータが存在しない

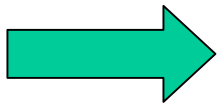


不完全データ

EM アルゴリズムにより欠損部の推定

親ノードの数と不完全データ

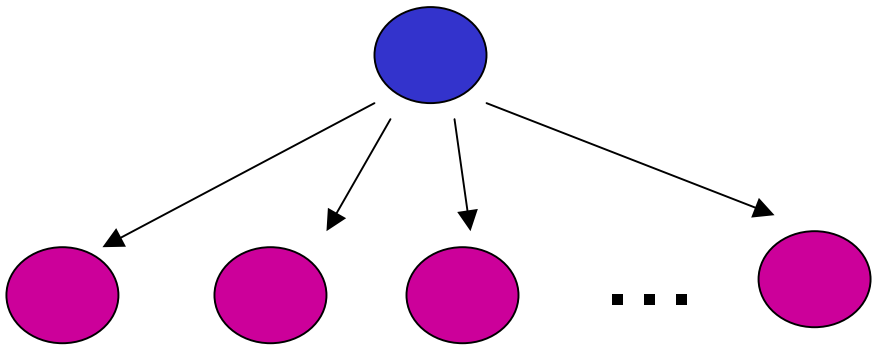
親ノードの数が多くなると CPT は指数的に大きくなる



不完全データが生じやすい

ナイーブベイズ

親ノードが1つ、CPT が小さく、実用的

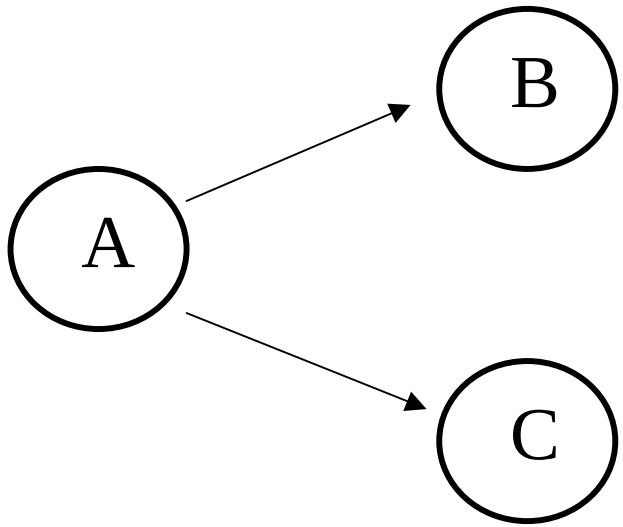


EM アルゴリズム（適用例 1/2）

不完全データ

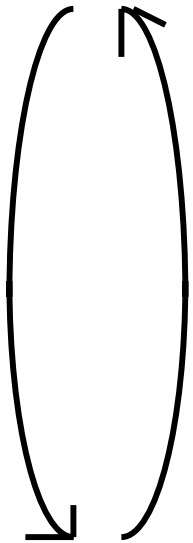
A	B	C
0	?	0
0	1	0
0	1	1
1	?	0
1	1	0
1	1	1
…	…	…

モデル

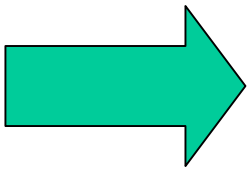


EM アルゴリズム（適用例 2/2）

E-step : 欠損データの推定



$$P(B=0 \mid A=0)=0.3 \qquad P(B=1 \mid A=0)=0.7$$
$$P(B=0 \mid A=1)=0.4 \qquad P(B=1 \mid A=1)=0.6$$

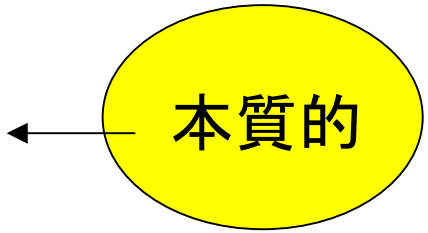


A	B	個数
0	1	2+0.7
0	0	0+0.3
1	0	0+0.4
1	1	2+0.6

集計表の更新

M-step : CPT の更新

更新された集計表から尤度を最大化



BN の EM学習

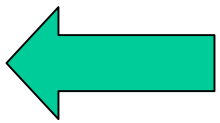
η ：初期ネットワーク

θ ：CPT の値

y ：欠損値

$\Omega(y)$ ：y を加えてできる完全データ

$$Q(\theta \mid \eta) = \sum_{x \in \Omega(y)} P(x \mid y, \eta) \log P(x \mid \theta)$$



最大化する θ を求める

（M-step)