欠損データがあっても最適なベイズ Chow-Liu アルゴリズム

鈴木譲

大阪大学

2012 年 10 月 16 日

Road Map

1. はじめに
2. ベイズ型 Chow-Liu アルゴリズム
3. 欠損データを含む場合の対応

おわび

PGM 2012

- 整数計画法を利用した BN の構造推定
  (今回は、成果に関する話がむずかしい)
- <u>欠損データがあっても最適なベイズ Chow-Liu アルゴリズム</u>

研究のねらい

機械学習への応用

例から、属性間の確率的依存関係を見出す

- Bayesian Network: 非巡回有向グラフによる表現
- Markov Network: 無向グラフによる表現

応用上の要請

- 属性数 N の多項式時間でないと、リアルタイムは無理
- EM とかではなく、不完全データの扱いを容易に扱えないか

Chow-Liu: 木への近似 (1968)

$X^{(1)}, \cdots, X^{(N)}$: N (≥ 1) 離散 確率変数
$P_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$: $X^{(1)} = x^{(1)}, \cdots, X^{(N)} = x^{(N)}$ の分布

$V := \{1, \cdots, N\}$ と $E \subseteq \{\{i,j\} | i \neq j, i,j \in V\}$ が木を構成することを仮定

$$Q_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)} | E) = \prod_{\{i,j\} \in E} \frac{P_{i,j}(x^{(i)}, x^{(j)})}{P_i(x^{(i)}) P_j(x^{(j)})} \prod_{i \in V} P_i(x^{(i)})$$

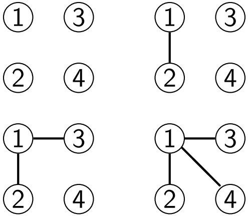
$D(P_{1,\cdots,N} || Q_{1,\cdots,N}) \to \min$
ループができない限り、相互情報量 $I(i,j)$ を最大にする $\{i,j\}$ を結ぶ

$$I(i,j) := \sum_{x^{(i)}, x^{(j)}} P_{i,j}(x^{(i)}, x^{(j)}) \log \frac{P_{i,j}(x^{(i)}, x^{(j)})}{P_i(x^{(i)}) P_j(x^{(j)})}$$

例

$I(i,j)$ を重みにもつ Kruscal のアルゴリズム

i	1	1	2	1	2	3
j	2	3	3	4	4	4
$I(i,j)$	12	10	8	6	4	2



Chow-Liu: 木を仮定した最尤推定

推定

分布 $P_{1,\cdots,N}$ ではなく、n 個の例 $x^n = \{(x_i^{(1)}, \cdots, x_i^{(N)})\}_{i=1}^n$ が利用可

x^n の相対頻度から得られる:

- $\hat{H}^n(x^n|E)$: 木を仮定した経験的エントロピー
- $\hat{I}(i,j)$: 木を仮定した $\{i,j\}$ の経験的相互情報量

$\hat{H}^n(x^n|E) \to \min$

ループができない限り、$\hat{I}(i,j)$ を最大にする $\{i,j\}$ を結ぶ

Chow-Liu: 木を仮定した Baye 推定 (Suzuki, 1993)

$$R^n(x^n|E) := \prod_{\{i,j\}\in E} \frac{R^n(i,j)}{R^n(i)R^n(j)} \prod_{i\in V} R^n(i)$$

$$R^n(i) := \frac{\Gamma(n+\alpha^{(i)}a)\Gamma(a)^{\alpha^{(i)}}}{\Gamma(\alpha^{(i)}a)\prod_{x^{(i)}}\Gamma(c_i[x^{(i)}]+a)}$$

$$R^n(i,j) := \frac{\Gamma(n+\alpha^{(i)}\alpha^{(j)}a)\Gamma(a)^{\alpha^{(i)}\alpha^{(j)}}}{\Gamma(\alpha^{(i)}\alpha^{(j)}a)\prod_{x^{(i)},x^{(j)}}\Gamma(c_{i,j}[x^{(i)},x^{(j)}]+a)}$$

($a>0$, $c_i[x^{(i)}], c_{i,j}[x^{(i)},x^{(j)}]$: 頻度, $\alpha^{(i)}$: $X^{(i)}$ が何個の値を取るか)

$$\sum_{\{x_k^{(i)}\}_{k=1}^n} R^n(i) = 1\ , \qquad \sum_{\{x_k^{(i)}\}_{k=1}^n,\{x_k^{(j)}\}_{k=1}^n} R^n(i,j) = 1$$

Chow-Liu: 木を仮定した Baye 推定 (Suzuki, 1993)

$\pi(E)$: 木 (V,E) の事前確率

$$J(i,j) := \frac{1}{n}\log\frac{R^n(i,j)}{R^n(i)R^n(j)}$$

$\pi(E)R^n(x^n|E) \to \max$ (π: 木の事前分布、一様とする)

ループができず、$J(i,j) \geq 0$ である限り、
$J(i,j)$ を最大にする $\{i,j\}$ を結ぶ

Chow-Liu: 木を仮定した MDL 推定 (Suzuki, 1993)

$a = 1/2$

$$L(x^n|E) := -\log R^n(x^n|E) \approx \hat{H}^n(x^n|E) + \frac{1}{2}k(E)\log n$$

$k(E)$: 木 (V, E) におけるパラメータの個数

$$J(i,j) \approx \hat{I}(i,j) - \frac{1}{2n}(\alpha^{(i)} - 1)(\alpha^{(j)} - 1)\log n$$

$\alpha^{(i)}$: $X^{(i)}$ が何個の値を取りうるか

$L(x^n|E) - \log \pi(E) \to \min$

ループができない限り、$J(i,j)$ を最大にする $\{i,j\}$ を結ぶ

ML と MDL

	ML	MDL
E の選択	$\hat{H}^n(x^n\|E)$ 最小	$\hat{H}^n(x^n\|E)+\frac{1}{2}k(E)\log n$ 最小
$\{i,j\}$ の選択	$\hat{I}(i,j)$ 最大	$\hat{I}(i,j)-\frac{1}{2n}(\alpha^{(i)}-1)(\alpha^{(j)}-1)\log n$ 最大
基準	x^n の E での適合性	x^n の E での適合性 E の簡潔性

欠損データを含む (一般的な) 場合の定式化

(V, E)

(木ではなく) 一般的な森、$\pi(E)$: 一定

従来 (欠損データなし): 最大化すべきもの

$$R^n(x^n|E) := \prod_{\{i,j\}\in E} \frac{R^n(i,j)}{R^n(i)R^n(j)} \prod_{i\in V} R^n(i)$$

今回 (欠損データがあってもよい): 最大化すべきもの

$$S^n(x^n|E) := \sum R^n(x^n|E)$$

$\sum$: 欠損している値のすべての場合について和

$S^n(x^n|E)$ の性質

$$S^n(x^n|E) := \sum R^n(x^n|E) = \prod_{\{i,j\}\in E} \frac{S^n(i,j)}{S^n(i)S^n(j)} \prod_{i\in V} S^n(i)$$

$$S^n(i) = \sum R^n(i)$$

$$S^n(i,j) = \sum R^n(i,j)$$

$S^n(i,j)$ の計算

$$S^n(i,j) = R^n_{i\cap j}(i,j) R^n_{i\setminus j}(i) R^n_{j\setminus i}(j)$$

$$x^n = \{x_k\}_{k=1}^n\ ,\ x_k = (x_k^{(1)}, \cdots, x_k^{(N)})$$

$$k = \underbrace{1, \cdots, r}_{X^{(i)}, X^{(j)}} \underbrace{r+1, \cdots, r+s}_{X^{(i)}} \underbrace{r+s+1, \cdots, r+s+t}_{X^{(j)}} \underbrace{r+s+t, \cdots, n}_{none}$$

$R^n_{i\cap j}(i,j)$: $\{x_k^{(i)}\}_{k=1}^r, \{x_k^{(j)}\}_{k=1}^r$ の予測確率
$R^n_{i\setminus j}(i)$: $\{x_k^{(i)}\}_{k=1}^r$ のもとでの $\{x_k^{(i)}\}_{k=r+1}^{r+s}$ の予測確率
$R^n_{j\setminus i}(j)$: $\{x_k^{(j)}\}_{k=1}^r$ のもとでの $\{x_k^{(j)}\}_{k=r+s+1}^{r+s+t}$ の予測確率

$S^n(i)$ の計算

$$S^n(i) = R^n_{i\cap j}(i) R^n_{i\setminus j}(i)$$

$$x^n = \{x_k\}_{k=1}^n \ , \ x_k = (x_k^{(1)}, \cdots, x_k^{(N)})$$

$$k = \underbrace{1, \cdots, r}_{X^{(i)}, X^{(j)}} \underbrace{r+1, \cdots, r+s}_{X^{(i)}} \underbrace{r+s+1, \cdots, r+s+t}_{X^{(j)}} \underbrace{r+s+t, \cdots, n}_{none}$$

$R^n_{i\cap j}(i)$: $\{x_k^{(i)}\}_{k=1}^r$ の予測確率
$R^n_{i\setminus j}(i)$: $\{x_k^{(i)}\}_{k=1}^r$ のもとでの $\{x_k^{(i)}\}_{k=r+1}^{r+s}$ の予測確率

$S^n(x^n|E)$ の計算

$$S^n(x^n|E) = \prod_{\{i,j\}\in E} \frac{S^n(i,j)}{S^n(i)S^n(j)} \prod_{i\in V} S^n(i) = \prod_{\{i,j\}\in E} \frac{R^n_{i\cap j}(i,j)}{R^n_{i\cap j}(i)R^n_{i\cap j}(j)} \prod_{i\in V} S^n(i)$$

$$J(i,j) := \frac{1}{n} \log \prod_{\{i,j\}\in E} \frac{R^n_{i\cap j}(i,j)}{R^n_{i\cap j}(i)R^n_{i\cap j}(j)}$$

結論

- データが完全の場合: $J(i,j)=\frac{1}{n}\log\frac{R^n(i,j)}{R^n(i)R^n(j)}$
- 一般の場合: $J(i,j)=\frac{1}{n}\log\frac{R^n_{i\cap j}(i,j)}{R^n_{i\cap j}(i)R^n_{i\cap j}(j)}$

ベイズ Chow-Liu アルゴリズム

- 欠損データがあっても、ベイズ最適な森を選択することが容易
- 同じ議論が、連続でも、離散と連続が混在でもできる (PGM 2012)