The Bayesian Chow-Liu Algorithm

Joe Suzuki

Osaka University

September 19, 2012
Granada, Spain

Chow-Liu: Tree Approximation (1968)

$X^{(1)}, \cdots, X^{(N)}$: N (≥ 1) discrete random variables
$P_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)})$: the distribution of $X^{(1)} = x^{(1)}, \cdots, X^{(N)} = x^{(N)}$

Assume $V := \{1, \cdots, N\}$ and $E \subseteq \{\{i,j\} | i \neq j, i, j \in V\}$ consist a tree.

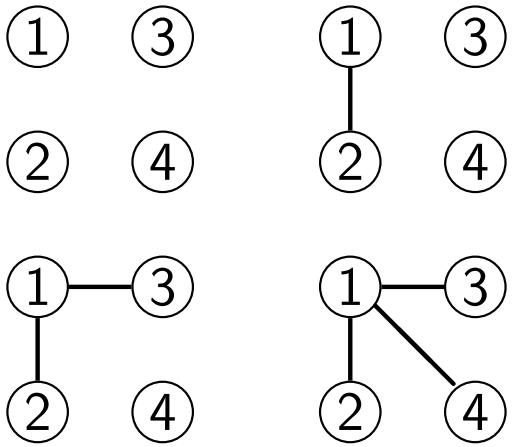
$$Q_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)}|E) = \prod_{\{i,j\}\in E} \frac{P_{i,j}(x^{(i)}, x^{(j)})}{P_i(x^{(i)})P_j(x^{(j)})} \prod_{i\in V} P_i(x^{(i)})$$

$D(P_{1,\cdots,N}||Q_{1,\cdots,N}) \rightarrow$ min

Connect $\{i,j\}$ with the largest $I(i,j)$ if no loop is generated

Example

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: Tree Estimation with ML

Estimation

Not $P_{1,\cdots,N}$ but n examples $x^n = \{(x_i^{(1)}, \cdots, x_i^{(N)})\}_{i=1}^n$ are available

$\hat{H}^n(x^n|E)$: the empirical entropy w.r.t. the tree obtained via the relative frequencies from x^n

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

Connect $\{i,j\}$ with the largest empirical $\hat{I}(i,j)$ $\cdots$

Chow-Liu: Tree Estimation with Bayes (Suzuki, 1993)

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

$\alpha^{(i)}$: how many values $X^{(i)}$ takes

$$R^n(i) := \frac{\Gamma(n+\alpha^{(i)}/2)\Gamma(a)^{\alpha^{(i)}}}{\Gamma(\alpha^{(i)}/2)\prod_{x^{(i)}}\Gamma(c_i[x^{(i)}]+1/2)}$$

$$R^n(i,j) := \frac{\Gamma(n+\alpha^{(i)}\alpha^{(j)}/2)\Gamma(1/2)^{\alpha^{(i)}\alpha^{(j)}}}{\Gamma(\alpha^{(i)}\alpha^{(j)}/2)\prod_{x^{(i)},x^{(j)}}\Gamma(c_{i,j}[x^{(i)},x^{(j)}]+1/2)}$$

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

$\pi(E)R^n(x^n|E) \to$ max (π: prior prob. assuming to be uniform)

Connect $\{i,j\}$ with the largest $J(i,j)\cdots$

Chow-Liu: Tree Estimation with MDL (Suzuki, 1993)

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

$k(E)$: # of parameters in the tree

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

$\alpha^{(i)}$: how many values $X^{(i)}$ takes

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

Connect $X^{(i)}, X^{(j)}$ with the largest $J(i,j)$ $\cdots$

ML vs MDL

	ML	MDL
selection of E	minimizes $\hat{H}^n(x^n\|E)$	minimizes $\hat{H}^n(x^n\|E)+\frac{1}{2}k(E)\log n$
selection of $\{i,j\}$	maxmizes $\hat{I}(i,j)$	maxmizes $\hat{I}(i,j)-\frac{1}{2n}(\alpha^{(i)}-1)(\alpha^{(j)}-1)\log n$
criterion	fitness of x^n to E	fitness of x^n to E simplicity of E

What if discrete and continuous variables are present

All the variables are discrete $\Longrightarrow$ unrealistic

In reality, some fields are discrete and others continuous in any database

What are

- Bayesian measures $R^n(i), R^n(j), R^n(i,j)$
- Bayesian estimator of mutual information

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

for the general case ?

Estimation of density functions

- $A_0 := \{A\}$ with $A := [0,1)$
- A_{j+1} is a refinement of A_j

$A_1 = \{[0,1/2),[1/2,1)\}$
$A_2 = \{[0,1/4),[1/4,1/2),[1/2,3/4),[3/4,1)\}$
$\cdots$
$A_j = \{[0,2^{-(j-1)}),[2^{-(j-1)},2\cdot 2^{-(j-1)}),\cdots,[(2^{j-1}-1)2^{-(j-1)},1)\}$
$\cdots$

Q_j^n: prediction probability w.r.t. A_j^n
$s_j : A \to A_j$ (quantization)
λ: Lebesgue measure (interval width)

$$g_j^n(x^n) := \frac{Q_j^n(s_j(x_1),\cdots,s_j(x_n))}{\lambda(s_j(x_1))\cdots\lambda(s_j(x_n))}\ ,\ x^n=(x_1,\cdots,x_n)\in A^n$$

$$\sum \omega_j = 1,\ \omega_j > 0,\ g^n(x^n) := \sum_j \omega_j g_j^n(x^n)$$

Ryabko 2009

$$f_j(x) := \frac{P(s_j(x))}{\lambda(s_j(x))} \text{ (density function for level } j\text{)}$$
$$f^n(x^n) := f(x_1)\cdots f(x_n)$$

Proposition

Suppose we choose $\{A_j\}$ s.t. $D(f||f_j) := E[\log \frac{f(X)}{f_j(X)}] \to 0$ as $j \to \infty$.

Then, for any f, as $n \to \infty$, a.e.

$$\frac{1}{n} \log \frac{f^n(x^n)}{g^n(x^n)} \to 0 \tag{1}$$

Estimation of generalized density functions

$B_0 := \{B\}$ with $B := \{1, 2, 3, \cdots\}$
$B_1 := \{\{1\}, \{2, 3, \cdots\}\}$
$B_2 := \{\{1\}, \{2\}, \{3, 4, \cdots\}\}$
$\cdots$
$B_k := \{\{1\}, \{2\}, \cdots, \{k\}, \{k+1, k+2, \cdots\}\}$
$\cdots$
Q_k^n: prediction probability w.r.t. B_k^n
$t_k : B \to B_k$ (quantization)

$$\eta(\{k\}) = \frac{1}{k} - \frac{1}{k+1}$$

$$g_k^n(y^n) := \frac{Q_k^n(t_k(y_1), \cdots, t_k(y_n))}{\eta(t_k(y_1)) \cdots \eta(t_k(y_n))}, \ y^n = (y_1, \cdots, y_n) \in B^n$$

$\sum \omega_k = 1$, $\omega_k > 0$, $g^n(y^n) := \sum_k \omega_k g_k^n(y^n)$

Suzuki 2011

$$f(y) := \frac{dP}{d\eta}(y),\ f_k(y) := \frac{P(s_k(y))}{\eta(s_k(y))}$$

Suppose that η is σ-finite, and that $P \ll \eta$.

Theorem 1 (estimation of generalized density functions)

Suppose we choose $\{B_k\}$ s.t.

$$D(f||f_k) := E[\log \frac{f(Y)}{f_k(Y)}] \to 0$$

as $k \to \infty$. Then, for any f, as $n \to \infty$, a.e.

$$\frac{1}{n} \log \frac{f^n(y^n)}{g^n(y^n)} \to 0 \tag{2}$$

$$(X, Y) \in A \times B$$

Q^n_{jk}: prediction probability w.r.t. $(A_j \times B_k)^n$

$$g^n_{jk}(x^n, y^n) := \frac{Q^n_{jk}(s_j(x_1), \cdots, s_j(x_n), t_k(y_1), \cdots, t_k(y_n))}{\lambda(s_j(x_1)) \cdots \lambda(s_j(x_n))\eta(t_k(y_1)) \cdots \eta(t_k(y_n))}$$

$\sum_{jk} \omega_{jk} = 1$, $\omega_{jk} > 0$, $g^n(x^n, y^n) := \sum_{j,k} \omega_{jk} g^n_{jk}(x^n, y^n)$

For any f, as $n \to \infty$, a.e.

$$\frac{1}{n} \log \frac{f^n(x^n, y^n)}{g^n(x^n, y^n)} \to 0 \tag{3}$$

Estimation of Mutual Information

Given $X^n = x^n$ and $Y^n = y^n$, from the strong law of large numbers:

$$\frac{1}{n}\log\frac{f^n(x^n,y^n)}{f^n(x^n)f^n(y^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i,y_i)}{f(x_i)f(y_i)} \to I(X,Y)$$

and (1)(2)(3), we obtain

Theorem 2

$$\frac{1}{n}\log\frac{g^n(x^n,y^n)}{g^n(x^n)g^n(y^n)} \to I(X,Y)$$

a.e. as $n \to \infty$

A Generalized Version of the Chow-Liu with Bayes/MDL

$R^n(x^n|E)$; a measure
$g^n(x^n|E)$: a generalized density function (contains R^n as a special case)

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

are replaced by the generalized version:

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

$g(x^n|E) \to \max$

Connect $X^{(i)}, X^{(j)}$ with the largest $J(i,j)$ $\cdots$

Computing $g^n(x^n)$: $O(nJ)$

$x^n = (x_1, \cdots, x_n)$

1. $g^n := 0$
2. for $j = 1, \cdots, J$
   1. $c[a] := 0$ for $a \in A_j$;
   2. $g_j^n := 1$;
   3. for $i = 1, \cdots, n$
      1. $a := s(x_i)$; // quantization
      2. $c[a] := c[a] + 1$;
      3. $g_j^n := g_j^n * \dfrac{c[a] + 1/2}{j + |A_j|/2} / \lambda(a)$;
   4. $g^n := g^n + w_j * g_j$;

Computing $g^n(x^n, y^n)$: $O(nJK)$

$x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$

1. $g^n := 0$
2. for $j = 1, \cdots, J$, $k = 1, \cdots, K$
   1. $c[a, b] := 0$ for $(a, b) \in A_j \times B_k$;
   2. $g^n_{jk} := 1$;
   3. for $i = 1, \cdots, n$
      1. $a := s_j(x_i)$; $b := t_k(y_i)$; // quantization
      2. $c[a, b] := c[a, b] + 1$;
      3. $g^n_{jk} := g^n_{jk} * \dfrac{c[a, b] + 1/2}{j + |A_j||B_k|/2} / \{\lambda(a)\eta(b)\}$;
   4. $g^n := g^n + w_{jk} * g_{jk}$;

Experiments (1)

$\{A_j\}_{j=1}^{J}, \{B_k\}_{k=1}^{K}$

$J = K$	n	$\frac{1}{n}\log\frac{f^n(x^n,y^n)}{g^n(x^n,y^n)}$	time (ms)
2	100	0.0307	1.23
2	1000	0.0281	10.67
4	100	0.0049	3.29
4	1000	0.0021	28.71

- The larger J, K, the better correctness
- Computation is linear with J, K and n

Experiments (2)

$\{A_j\}_{j=1}^{J}$ with $J = 4$ and $n = 1000$

$f = f_2$	
j	$g_j^n(x^n)$
1	0.307
2	0.981
3	0.198
4	0.097

$f = f_4$	
j	$g_j^n(x^n)$
1	0.083
2	0.141
3	0.198
4	0.797

$$g^n(x^n) = \sum_j w_j g_j(x^n) \text{ vs } \max_j w_j g_j(x^n)$$

Conclusion

- Reformulate (Suzuki 1993) as the Bayes Chow-Liu
- Apply Bayes measures without assuming either discrete or continuous (Suzuki 2011)
- Propose Bayesian mutual information estimator

Merits

- easy to execute (could be R commands)
- easy to embed prior information (a merit of Bayes)
- mixure of quantizations: more robust than selecting one quantization

Other Applictions:

- Bayesian network structure estimation
- Markov order estimation (either discrete or continuous)