Parameter learning with EM for discrete BNs

Pablo M. Olmos, olmos@tsc.uc3m.es

Course on Bayesian Networks, November 2016

Index

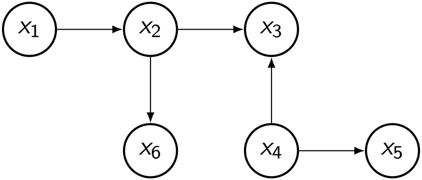
Section 1

Learning parameters with full observations

Motivation

- Let $X_j \in \mathcal{X}$, $j = 1, \ldots, 5$, be discrete R.V., where $K \doteq |\mathcal{X}|$
- Consider the following BN:

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_4)p(x_3|x_2, x_4)p(x_5|x_4)p(x_6|x_2)$$



- **BN structure is known**, **CPD tables are unknown**.
- Our goal is to **estimate the CPD tables from N independent samples** $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, drawn from $p(\boldsymbol{x})$.

Log-Likelihood of data (I)

- Let $\boldsymbol{x} \in \mathcal{X}^V$ be a discrete R.V. such that $p(\boldsymbol{x}) = \prod_{t=1}^{V} p(x_t|\boldsymbol{x}_{\mathrm{pa}(t)})$.
- Consider N independent samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}$. For each sample, we can write each $p(x_t|\boldsymbol{x}_{\mathrm{pa}(t)})$ term, $t = 1, \ldots, V$ as follows:

$$p(x_{it}|\boldsymbol{x}_{i,\mathrm{pa}(t)}) = \prod_{c=1}^{K_{\mathrm{pa}(t)}} \prod_{k=1}^{K_t} \theta_{tck}^{\mathbb{1}[x_{it}=k, \boldsymbol{x}_{i,\mathrm{pa}(t)}=c]}$$

- $\prod_{c=1}^{K_{\mathbf{pa}(t)}}$ → product over all possible values of $\boldsymbol{x}_{i,\mathrm{pa}(t)}$.
- $\prod_{k=1}^{K}$ → product over all possible values of $x_{i,t}$.
- $\theta_{tck}^{\mathbb{1}[x_{it}=k, \boldsymbol{x}_{i,\mathbf{pa}(t)}=c]}$ → equal to θ_{tck} if $x_{it} = k$, and $\boldsymbol{x}_{i,\mathrm{pa}(t)} = c$. Thus,

$$p(x_t = k|\boldsymbol{x}_{\mathrm{pa}(t)} = c) \doteq \theta_{tck}$$

- Note that $\sum_{k=1}^{K_t} \theta_{tck} = 1$.

Log-Likelihood of data (II)

- The log-likelihood of the complete data is given by

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{t=1}^{V} \sum_{c=1}^{K_{\mathrm{pa}(t)}} \sum_{k=1}^{K_t} N_{tck} \log \theta_{tck}$$

  where

$$N_{tck} = \sum_{i=1}^{N} \mathbb{1}[x_{it} = k, \boldsymbol{x}_{i,\mathrm{pa}(t)} = c]$$

  are the empirical counts.

- Log-likelihood is maximized if we take

$$\hat{\theta}_{tck} = \frac{N_{tck}}{\sum_{k'=1}^{K_t} N_{tck'}}$$

- **ML solution is very simple! We calculate frequencies!**

uc3m

MAP solution

- Dirichlet prior over $\boldsymbol{\theta}_{tc} = [\theta_{tc1}, \theta_{tc2}, \ldots, \theta_{tcK_t}]$:

$$\boldsymbol{\theta}_{tc} \sim \mathrm{Dir}\left[\alpha_{tc1}, \alpha_{tc2}, \ldots, \alpha_{tcK_t}\right]$$

- It is easy to show that $p(\boldsymbol{\theta}_{tc}|\mathcal{D})$ is another Dirichlet distribution with parameters

$$\boldsymbol{\theta}_{tc} \sim \mathrm{Dir}\left[\alpha_{tc1} + N_{tc1}, \alpha_{tc2} + N_{tc2}, \ldots, \alpha_{tcK_t} + N_{tcK_t}\right]$$

- The mean of θ w.r.t. the posterior distribution is

$$\mathbb{E}_{p(\boldsymbol{\theta}_{tc}|\mathcal{D})}[\theta_{tck}] = \frac{\alpha_{tck} + N_{tck}}{\sum_{k'=1}^{K_t} \alpha_{tck'} + N_{tck'}}$$

- θ_{tck} act as *pseudocounts*, avoiding to assign zero probability to unobserved outcomes in our data

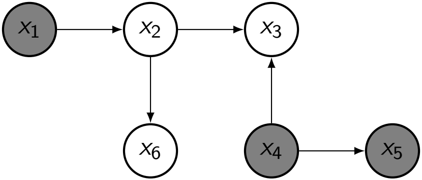
Section 2

Learning with partial observations

Motivation

- Let $X_j \in \mathcal{X}$, $j = 1, \ldots, 5$, be discrete R.V., where $K \doteq |\mathcal{X}|$
- Consider the following BN:

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_4)p(x_3|x_2, x_4)p(x_5|x_4)p(x_6|x_2)$$



- **BN structure is known**, **CPD tables are unknown**.
- Our goal is to **estimate the CPD tables from N independent samples** $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, drawn from $p(\boldsymbol{x})$.
- Only a few elements of $\boldsymbol{x}$ are observed!

EM in a nutshell

- $p(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{\theta})$, where $\boldsymbol{y}_i$ is observed and $\boldsymbol{x}_i$ is hidden, $i = 1, \ldots, N$.
- Our goal is to estimate $\boldsymbol{\theta}$ to maximize $p(\mathcal{D}|\boldsymbol{\theta})$ (ML) or as the mode of the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. **Complex! We have to marginalize first over $\boldsymbol{x}_i$, $i = 1, \ldots, N$.**
- EM algorithm. Initialize $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^0$. For $\ell = 1, 2, \ldots$
  1. **E-step**:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) &= \sum_{i=1}^{N} \int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{\ell-1}) \log\left(p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{x}_i \\
&= \sum_{i=1}^{N} \mathbb{E}_{p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{\ell-1})}[\log\left(p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{\theta})\right)]
\end{aligned}
$$

  2. **M-step**:

$$
\boldsymbol{\theta}^{\ell} = \arg\max_{\boldsymbol{\theta}}\ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) \qquad \text{(ML estimation)}
$$

$$
\boldsymbol{\theta}^{\ell} = \arg\max_{\boldsymbol{\theta}}\ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) + \log p(\boldsymbol{\theta}) \qquad \text{(MAP estimation)}
$$

EM for discrete BNs (I)

- $\boldsymbol{y}_i \to$ Set of observed variables for i-th data, $i = 1, \ldots, N$.
- The log-likelihood of the complete data is given by

$$\log p(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^{N} \| \boldsymbol{\theta}) = \sum_{t=1}^{V} \sum_{c=1}^{K_{\text{pa}(t)}} \sum_{k=1}^{K_t} N_{tck} \log \theta_{tck}$$

where

$$N_{tck} = \sum_{i=1}^{N} \mathbb{1}[x_{it} = k, \boldsymbol{x}_{i,\text{pa}(t)} = c]$$

uc3m

EM for discrete BNs (II)

- **E-step**:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) = \sum_{i=1}^{N} \sum_{c=1}^{K_{\text{pa}(t)}} \sum_{k=1}^{K_t} \widetilde{N_{tck}} \log \theta_{tck}$$

where

$$\begin{aligned} \widetilde{N_{tck}} &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}[x_{it} = k, \boldsymbol{x}_{i,\text{pa}(t)} = c]\right] \\ &= \sum_{i=1}^{N} p(x_{it} = k, \boldsymbol{x}_{i,\text{pa}(t)} = c | \boldsymbol{y}_i, \boldsymbol{\theta}^{\ell-1}) \end{aligned}$$

- Thus, given each pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, we simply have to compute the marginal joint probabilities $p(x_{it} = k, \boldsymbol{x}_{i,\text{pa}(t)} = c | \boldsymbol{y}_i)$, $t = 1, \ldots, V$.
- We will use **Belief Propagation for this task!**

EM for discrete BNs (III)

- **M-step**:

$$\hat{\theta}_{tck} = \frac{\widetilde{N_{tck}}}{\sum_{k'=1}^{K_t} \widetilde{N_{tck'}}} \qquad \text{(ML estimation)}$$

$$\hat{\theta}_{tck} = \frac{\alpha_{tck} + \widetilde{N_{tck}}}{\sum_{k'=1}^{K_t} \alpha_{tck'} + \widetilde{N_{tck'}}} \qquad \text{(MAP estimation)}$$