

732A96/TDDE15 Advanced Machine Learning

Graphical Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 3: Parameter Learning

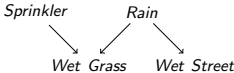
Contents

- ▶ Parameter Learning for BNs
 - ▶ Maximum Likelihood
- ▶ Parameter Learning for MNs
 - ▶ Iterative Proportional Fitting Procedure

Literature

- ▶ Main source
 - ▶ Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapter 8.
- ▶ Additional source
 - ▶ Koski, T. J. T. and Noble, J. M. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda* 40, 51-103, 2012.

Parameter Learning for BNs: Maximum Likelihood

DAG	Parameter values for the conditional probability distributions
 <pre> graph TD Sprinkler --> WetGrass[Wet Grass] Rain --> WetGrass Rain --> WetStreet[Wet Street] </pre>	$q(s) = (0.3, 0.7) = (\theta_{s_0}, \theta_{s_1})$ $q(r) = (0.5, 0.5) = (\theta_{r_0}, \theta_{r_1})$ $q(wg r_0, s_0) = (0.1, 0.9) = (\theta_{wg_0 r_0, s_0}, \theta_{wg_1 r_0, s_0})$ $q(wg r_0, s_1) = (0.7, 0.3) = (\theta_{wg_0 r_0, s_1}, \theta_{wg_1 r_0, s_1})$ $q(wg r_1, s_0) = (0.8, 0.2) = (\theta_{wg_0 r_1, s_0}, \theta_{wg_1 r_1, s_0})$ $q(wg r_1, s_1) = (0.9, 0.1) = (\theta_{wg_0 r_1, s_1}, \theta_{wg_1 r_1, s_1})$ $q(ws r_0) = (0.1, 0.9) = (\theta_{ws_0 r_0}, \theta_{ws_1 r_0})$ $q(ws r_1) = (0.7, 0.3) = (\theta_{ws_0 r_1}, \theta_{ws_1 r_1})$ $p(s, r, wg, ws) = q(s)q(r)q(wg s, r)q(ws r)$

- In general,

$$q(X_i = k | Pa_i = j) = \theta_{X_i=k | Pa_i=j}$$

- Recall that

$$p(X_i = k | Pa_i = j) = q(X_i = k | Pa_i = j)$$

Parameter Learning for BNs: Maximum Likelihood

- Given a sample $d_{1:N}$, the log likelihood function is

$$\begin{aligned}\log p(d_{1:N}|\theta, G) &= \log \prod_i p(d_i|\theta, G) = \log \prod_i \prod_{a_i} p(d_i[X_i]|d_i[Pa_i], \theta) \\&= \log \prod_i \prod_{a_i} \theta_{X_i=d_i[X_i]|Pa_i=d_i[Pa_i]} = \log \prod_i \prod_j \prod_k \theta_{X_i=k|Pa_i=j}^{N_{ijk}} \\&= \sum_i \sum_j \sum_k N_{ijk} \log \theta_{X_i=k|Pa_i=j}\end{aligned}$$

where N_{ijk} is the number of instances in $d_{1:N}$ with $X_i = k$ and $Pa_i = j$.

- To maximize the log likelihood function subject to the constraint $\sum_k \theta_{X_i=k|Pa_i=j} = 1$ for all i and j , we maximize

$$\sum_i \sum_j \sum_k N_{ijk} \log \theta_{X_i=k|Pa_i=j} + \sum_i \sum_j \lambda_{ij} (\sum_k \theta_{X_i=k|Pa_i=j} - 1)$$

where λ_{ij} are called Lagrange multipliers.¹

- Setting to zero the derivative with respect to $\theta_{X_i=k|Pa_i=j}$ gives

$$\theta_{X_i=k|Pa_i=j} = -N_{ijk}/\lambda_{ij}$$

- Replacing in the constraint gives $\lambda_{ij} = -N_{ij}$ and $\theta_{X_i=k|Pa_i=j}^{ML} = N_{ijk}/N_{ij}$.

¹Any stationary point of the Lagrangian function is a stationary point of the original function subject to the constraints. Moreover, the log likelihood function is concave.

Parameter Learning for MNs: Iterative Proportional Fitting Procedure

- Given a complete sample $d_{1:N}$, the log likelihood function is

$$\log p(d_{1:N}|\theta, G) = \sum_{K \in Cl(G)} \sum_k N_k \log \varphi(k) - N \log Z$$

where N_k is the number of instances in $d_{1:N}$ with $K = k$. Then

$$\log p(d_{1:N}|\theta, G)/N = \sum_{K \in Cl(G)} \sum_k p_e(k) \log \varphi(k) - \log Z$$

where $p_e(X)$ is the empirical probability distribution obtained from $d_{1:N}$.

- Let $Q \in Cl(G)$. The derivative with respect to $\varphi(q)$ is

$$\frac{\partial \log p(d_{1:N}|\theta, G)/N}{\partial \varphi(q)} = \frac{p_e(q)}{\varphi(q)} - \frac{1}{Z} \frac{\partial Z}{\partial \varphi(q)}$$

- Let $Y = X \setminus Q$. Then

$$\frac{\partial Z}{\partial \varphi(q)} = \sum_y \prod_{K \in Cl(G) \setminus Q} \varphi(k, \bar{k}) = \frac{Z}{\varphi(q)} \sum_y \prod_{K \in Cl(G) \setminus Q} \varphi(k, \bar{k}) \frac{\varphi(q)}{Z} = \frac{Z}{\varphi(q)} p(q|\theta, G)$$

where \bar{k} denotes the elements of q corresponding to the elements of $K \cap Q$.

- Putting together the results above, we have that

$$\frac{\partial \log p(d_{1:N}|\theta, G)/N}{\partial \varphi(q)} = \frac{p_e(q)}{\varphi(q)} - \frac{p(q|\theta, G)}{\varphi(q)}$$

Parameter Learning for MNs: Iterative Proportional Fitting Procedure

- ▶ Setting the derivative to zero gives ²

$$\varphi^{ML}(q) = \varphi(q)p_e(q)/p(q|\theta, G)$$

No closed form solution but ...

IPFP

Initialize $\varphi(k)$ for all $K \in Cl(G)$

Repeat until convergence

Set $\varphi(k) = \varphi(k)p_e(k)/p(k|\theta, G)$ for all $K \in Cl(G)$

- ▶ IPFP increases $\log p(d_{1:N}|\theta, G)$ in each iteration. So, it is globally optimal.
- ▶ Iterative coordinate ascend method.
- ▶ Note that computing $p(k|\theta, G)$ in the last line requires inference. Moreover, the multiplication and division are elementwise.
- ▶ Note also that Z needs to be computed in each iteration, which is computationally hard. This can be avoided by a careful initialization.

²The log likelihood function is concave.

Contents

- ▶ Parameter Learning for BNs
 - ▶ Maximum Likelihood
- ▶ Parameter Learning for MNs
 - ▶ Iterative Proportional Fitting Procedure

Thank you