Machine learning: lecture 24

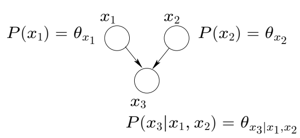
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Outline

- Learning Bayesian networks: complete data
  - estimating the parameters with fixed structure
  - learning the graph structure
- Learning Bayesian networks: incomplete data
  - EM and structural EM
- Review

Probabilities and conditional tables

- Simple example: three discrete variables, each taking m possible values



We assume that the model is fully parameterized in the sense that $\{\theta_{x_1}\}$, $\{\theta_{x_2}\}$, and $\{\theta_{x_3|x_1,x_2}$ for each distinct configuration of x_1 and $x_2\}$ are unrestricted and can be chosen independently of each other.

Likelihood and complete data

- When the observed data points are complete, the likelihood has a simple form:

$$\begin{aligned} P(D|G,\theta) &= \prod_{t=1}^{n} P(x_1^t)P(x_2^t)P(x_3^t|x_1^t,x_2^t) \\ &= \left(\prod_{x_1} P(x_1)^{N(x_1)}\right) \times \left(\prod_{x_2} P(x_2)^{N(x_2)}\right) \\ &\quad \times \prod_{x_1,x_2}\left(\prod_{x_3} P(x_3|x_1,x_2)^{N(x_1,x_2,x_3)}\right) \end{aligned}$$

Likelihood and complete data

- When the observed data points are complete, the likelihood has a simple form:

$$\begin{aligned} P(D|G,\theta) &= \prod_{t=1}^{n} P(x_1^t)P(x_2^t)P(x_3^t|x_1^t,x_2^t) \\ &= \left(\prod_{x_1} \theta_{x_1}^{N(x_1)}\right) \times \left(\prod_{x_2} \theta_{x_2}^{N(x_2)}\right) \\ &\quad \times \prod_{x_1,x_2}\left(\prod_{x_3} \theta_{x_3|x_1,x_2}^{N(x_1,x_2,x_3)}\right) \end{aligned}$$

Each conditional table such as $\theta_{x_3|x_1,x_2}$ for a fixed x_1 and x_2, can be estimated separately based on the observed counts $N(x_1, x_2, x_3)$.

ML parameter estimates

$$\begin{aligned} P(D|G,\theta) &= \left(\prod_{x_1} \theta_{x_1}^{N(x_1)}\right) \times \left(\prod_{x_2} \theta_{x_2}^{N(x_2)}\right) \\ &\quad \times \prod_{x_1,x_2}\left(\prod_{x_3} \theta_{x_3|x_1,x_2}^{N(x_1,x_2,x_3)}\right) \end{aligned}$$

- The maximum likelihood estimates of parameters $\theta_{\cdot|x_1,x_2} = \{\theta_{1|x_1,x_2}, \ldots, \theta_{m|x_1,x_2}\}$ for each fixed configuration of x_1 and x_2 are simply normalized counts (cf. Markov models):

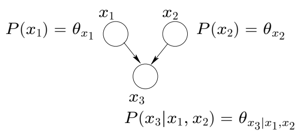
$$\hat{\theta}_{x_3|x_1,x_2} = \frac{N(x_1,x_2,x_3)}{N(x_1,x_2)}$$

where $N(x_1,x_2) = \sum_{x_3} N(x_1,x_2,x_3)$.

Bayesian estimates: the prior

- We introduce independent priors, e.g., $P_3(\theta_{\cdot|x_1,x_2})$, across variables and for each distinct configuration of the parents

$$P(\theta|G) = P_1(\theta_{\cdot}) \times P_2(\theta_{\cdot}) \times \prod_{x_1,x_2} P_3(\theta_{\cdot|x_1,x_2})$$



Bayesian estimates: the prior

- We introduce independent priors, e.g., $P_3(\theta_{\cdot|x_1,x_2})$, across variables and for each distinct configuration of the parents
- Moreover, we assume that these priors are Dirichlet:

$$P_3(\theta_{\cdot|x_1,x_2}) = \frac{1}{Z'} \prod_{x_3} \theta_{x_3|x_1,x_2}^{N'(x_1,x_2,x_3)-1}$$

with hyper-parameters (parameters of the prior) $N'(x_1,x_2,x_3) \geq 0$, interpreted as prior "counts".

Bayesian estimates: the prior

- We introduce independent priors, e.g., $P_3(\theta_{\cdot|x_1,x_2})$, across variables and for each distinct configuration of the parents
- Moreover, we assume that these priors are Dirichlet:

$$P_3(\theta_{\cdot|x_1,x_2}) = \frac{1}{Z'} \prod_{x_3} \theta_{x_3|x_1,x_2}^{N'(x_1,x_2,x_3)-1}$$

with hyper-parameters (parameters of the prior) $N'(x_1,x_2,x_3) \geq 0$, interpreted as prior "counts".

This prior is concentrated around

$$\theta'_{x_3|x_1,x_2} = \frac{N'(x_1,x_2,x_3)}{N'(x_1,x_2)}$$

where $N'(x_1,x_2) = \sum_{x_3} N'(x_1,x_2,x_3)$, and more so for larger values of $N'(x_1,x_2)$.

Bayesian estimates: the posterior

- The posterior is also Dirichlet

$$P_3(\theta_{\cdot|x_1,x_2}|D) \propto \overbrace{\prod_{x_3} \theta_{x_3|x_1,x_2}^{N(x_1,x_2,x_3)}}^{\text{likelihood}} \cdot \overbrace{\prod_{x_3} \theta_{x_3|x_1,x_2}^{N'(x_1,x_2,x_3)-1}}^{\text{prior}}$$

$$= \frac{1}{Z} \prod_{x_3} \theta_{x_3|x_1,x_2}^{N'(x_1,x_2,x_3)+N(x_1,x_2,x_3)-1}$$

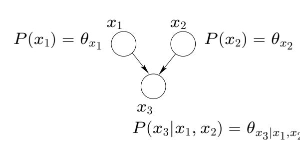
with hyper-parameters $N'(x_1,x_2,x_3) + N(x_1,x_2,x_3)$.

Outline

- Learning Bayesian networks: complete data
  - estimating the parameters with fixed structure
  - learning the graph structure
- Learning Bayesian networks: incomplete data
  - EM and structural EM
- Review

Bayesian score

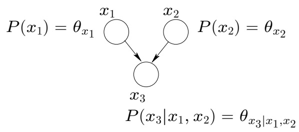
- We can use the marginal likelihood (Bayesian score) as a model (graph) selection criterion:



$$P(D|G) = \int P(D|G,\theta)P(\theta|G)d\theta$$

Bayesian score

- We can use the marginal likelihood (Bayesian score) as a model (graph) selection criterion:



$$P(D|G) = \int P(D|G,\theta)P(\theta|G)d\theta$$

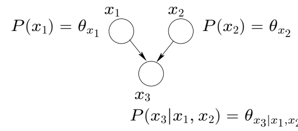
- The form of the likelihood and the prior

$$P(D|G,\theta) = \prod_{x_1} \theta_{x_1}^{N(x_1)} \times \prod_{x_2} \theta_{x_2}^{N(x_2)} \times \prod_{x_1,x_2} \prod_{x_3} \theta_{x_3|x_1,x_2}^{N(x_1,x_2,x_3)}$$

$$P(\theta|G) \propto \underbrace{\prod_{x_1} \theta_{x_1}^{N'(x_1)}}_{P_1(\theta_\cdot)} \times \underbrace{\prod_{x_2} \theta_{x_2}^{N'(x_2)}}_{P_2(\theta_\cdot)} \times \prod_{x_1,x_2} \underbrace{\prod_{x_3} \theta_{x_3|x_1,x_2}^{N'(x_1,x_2,x_3)}}_{P_3(\theta_{\cdot|x_1,x_2})}$$

permit us to evaluate the Bayesian score locally.

Bayesian score: graphs



- The Bayesian score reduces to a product of local terms:

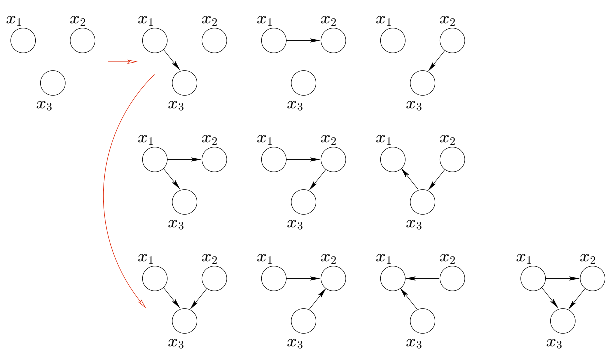
$$P(D|G) = \int P(D|G,\theta)P(\theta|G)d\theta$$

$$= \int \left[\prod_{x_1} \theta_{x_1}^{N(x_1)} \times \prod_{x_2} \theta_{x_2}^{N(x_2)} \times \prod_{x_1,x_2} \prod_{x_3} \theta_{x_3|x_1,x_2}^{N(x_1,x_2,x_3)} \right] P(\theta|G)d\theta$$

$$= P(D_1|G) \times P(D_2|G) \times \prod_{x_1,x_2} P(D_{3|x_1,x_2}|G)$$

Learning Bayesian networks

- We can perform a greedy search over (equivalence classes of) Bayesian networks based on the score:



Review for the final

- The final is comprehensive
- Major concepts
  - regression, active learning
  - classification, margins, kernels, feature selection
  - over-fitting, regularization, generalization, model selection
  - latent variable models, estimation with incomplete data
  - clustering, objectives
  - graphs and probabilities, inference