208891 Probabilistic Graphical Models **Spring 2019**

Lecture 6: January 13

Lecturer: Donlapark Pornnopparath

6.1 Bayesian parameter estimation

Suppose that we have a simple network $X \to Y$ where X is controlled by a parameter θ_X and Y is controlled by a parameter $\theta_{Y|X}$. We assume that the parameters are not shared i.e. They are independent. To see that the parameters are independent a posteriori, we note that any path between θ_X and $\theta_{Y|X}$ has the form

$$\theta_X \to X_i \to Y_i \leftarrow \theta_{Y|X},$$

for $i = 1, 2, \ldots, n$. Since the node X_i blocks the path, θ_X and $\theta_{Y|X}$ are d-separated given data D, which implies

$$p(\theta_X, \theta_{Y|X}|D) = p(\theta_X|D)p(\theta_{Y|X}|D). \tag{6.1}$$

This allows us to solve for the posterior distribution separately: for example, if each variable follows the Dirichlet-multinomial model, then the joint posterior over both parameters is the product of Dirichlet posterior of each parameter. Suppose that $D = \{x_1, y_1, \ldots, x_n, y_n\}$ and we want to make a prediction for (x_{n+1}, y_{n+1}). The joint posterior is given by

$$\begin{aligned}
p(x_{n+1}, y_{n+1}|D) &= \int p(x_{n+1}, y_{n+1}|D, \theta)p(\theta|D)\ d\theta \\
&= \int\int p(x_{n+1}, y_{n+1}|\theta)p(\theta_X|D)p(\theta_{Y|X}|D)\ d\theta_X d\theta_{Y|X} \\
&= \int\int p(x_{n+1}|\theta_X)p(y_{n+1}|x_{n+1}, \theta_{Y|X}) \\
&\quad\ p(\theta_X|D)p(\theta_{Y|X}|D)\ d\theta_X d\theta_{Y|X} \\
&= \left(\int p(x_{n+1}|\theta_X)p(\theta_X|D)\ d\theta_X\right) \\
&\quad\ \left(\int p(y_{n+1}|x_{n+1}, \theta_{Y|X})p(\theta_{Y|X}|D)\ d\theta_{Y|X}\right).
\end{aligned}$$

Now we focus on a more general graph G. The posterior distribution given the data D is

$$p(\theta|D) = \frac{L(\theta|D)p(\theta)}{p(D)}.$$

As we have mentioned before, the likelihood can be decomposed as

$$L(\theta|D) = \prod_i L_i(\theta_{X^i|\mathrm{pa}_{X^i}}|D),$$

and we assume that all parameters are independent

$$p(\theta) = \prod_i p(\theta_{X^i|\mathrm{pa}_{X^i}}).$$

$p(D)$ is the normalizing constant that we will discuss later. Combining these together, we see that the posterior can be represented as a product of local terms:

$$p(\theta|D) = \frac{1}{p(D)} \prod_i \left[L_i(\theta_{X^i|\text{pa}_{X^i}}|D) p(\theta_{X^i|\text{pa}_{X^i}}) \right]$$

Similarly to the two-variable network, the probability of obtaining $x^1_{n+1}, \ldots, x^m_{n+1}$ as the next instance is

$$\begin{aligned} &p(x^1_{n+1}, \ldots, x^m_{n+1}|D) \\ &\quad = \prod_i \int p(x^i_{n+1}|\text{pa}_{x^i,n+1}, \theta_{X^i|\text{pa}_{X^i}}) p(\theta_{X^i|\text{pa}_{X^i}}|D) \, d\theta_{X^i|\text{pa}_{X^i}}, \end{aligned}$$

where $\text{Pa}_{x^i,m+1}$ denotes all parent variables of x^i in the $n+1$–th instance.

Example 6.1. Let us go back to the two-variable network where $X \in \{x^0, x^1\}$ and $Y \in \{y^0, y^1\}$ are discrete variables. In this case, a suitable choice of prior is Dirichlet for both θ_X and $\theta_{Y|X}$. From the decomposition (6.1), we have to compute

$$p(\theta_X|D) \propto L(\theta_X|D) p(\theta_X),$$

which is straightforward to compute. We have to be careful about the other term $p(\theta_{Y|X}$ since the conditional independence between $\theta_{Y|x^0}$ and $\theta_{Y|x^1}$ is not clear. To see this, we observe that the path from $\theta_{Y|x^1}$ to Y_i disappears when we observe $X_i = x^0$ and the path from $\theta_{Y|x^0}$ to Y_i disappears when we observe $X_i = x^1$. Since this is true for all i, we see that $\theta_{Y|x^0}$ and $\theta_{Y|x^1}$ are conditional independent given D. Therefore, we have

$$\begin{aligned} p(\theta_{Y|X}|D) &\propto L(\theta_{Y|X}|D) p(\theta_{Y|X}) \\ &= L(\theta_{Y|x^0}|D) p(\theta_{Y|x^0}) L(\theta_{Y|x^1}|D) p(\theta_{Y|x^1}) \\ &= p(\theta_{Y|x^0}) \prod_{i:x_i=x^0} p(y_i|x_i, \theta_{Y|x^0}, \theta_{Y|x^1}) \\ &\quad p(\theta_{Y|x^1}) \prod_{i:x_i=x^1} p(y_i|x_i, \theta_{Y|x^1}, \theta_{Y|x^1}) \\ &= p(\theta_{Y|x^0}) \prod_{i:x_i=x^0} p(y_i|x_i, \theta_{Y|x^0}) \\ &\quad p(\theta_{Y|x^1}) \prod_{i:x_i=x^1} p(y_i|x_i, \theta_{Y|x^1}). \end{aligned}$$

This is a product of separate joint distribution of Dirichlet prior. Suppose that $\theta_{Y|x^0} \sim \text{Dir}(\alpha_{y^0|x^0}, \alpha_{y^1|x^0})$. Using the rule for updating a single Dirichlet yields

$$\theta_{Y|x^0,D} \sim \text{Dir}\left(\alpha_{y^0|x^0} + n(x^0, y^0), \alpha_{y^1|x^1} + n(x^1, y^1)\right).$$

◇

In general, when we have a node $X \in \{x^1, x^2, \ldots, x^K\}$, its parents $U = \text{Pa}_X$ and Dirichlet prior $\alpha_{x^1|u}, \ldots, \alpha_{x^K|u}$, then the posterior distribution is

$$\theta_{X|u,D} \sim \text{Dir}\left(\alpha_{x^1|u} + n(x^1, u), \alpha_{x^K|u} + n(x^K, u)\right).$$

Thus the probability of a new observation $X^i_{n+1} = x$ given that $U_{n+1} = u$ is

$$\mathbb{P}(X^i_{n+1} = x|U_{n+1} = u, D) = \frac{\alpha_{x|u} + n(x, u)}{\sum_i \alpha_{x^i|u} + n(x^i, u)}$$

6.2 Choosing parameter priors

Given a node X^i which has a set of distributions $\text{Mul}(\theta_{X^i|\text{pa}_{X^i}})$, one for each possible value of Pa_{X^i}. Each of the parameters has a Dirichlet distribution with hyperparameters

$$\alpha_{X^i|\text{pa}_{X^i}} = (\alpha_{x^1|\text{pa}_{x^1}}, \ldots, \alpha_{X^K|\text{pa}_{X^K}})$$

There are some suggested way of specifying these hyperparameters:

- Assign values to these hyperparameters based on expert knowledge.
- Use a fixed prior e.g. $\alpha_{x^j|\text{pa}_{x^1}} = 1$ for all hyperparameters.
- Construct an "imaginary dataset"i D' from our prior experience and set

  $$\alpha_{x|\text{pa}_x} = n(x, \text{Pa}_x),$$

  where $n(x, \text{Pa}_x)$ is the number of instances that $X = x$ and $Pa_x = pa_x$.