208891 Probabilistic Graphical Models **Spring 2019**

Lecture 4: January 6

Lecturer: Donlapark Pornnopparath

In the next few lectures, we will focus on structure learning of a Bayesian network. One way to find a suitable graph given some observed data D is to search, among all possible graphs G, for the one that maximizes the score metric:

$$Score(G : D) = LL(G : D) - \phi(|D|)\|G\|,$$

where LL is the log-likelihood of the data under the graph structure G with its parameters estimated using MLE, $|D|$ is the number of samples and $\|G\|$ is the number of parameters in G. The second term is for penalizing an overfitting model, such as a complete graph. We notice that the metric is the AIC when $\phi(t) = 1$ and BIC when $\phi(t) = \log(t)/2$. To compute the first term, we need to figure out how to make MLE inference based on a Bayesian graph G.

4.1 Maximum likelihood estimation

We begin with a basic example of parameter estimation based on MLE.

Example 4.1. Suppose that we are tossing a coin whose probability of turning head is θ. Then the distribution of this event is $\text{Ber}(\theta)$. Assume that after five independent tosses, we obtain the following outcome:

$$H, T, T, H, H.$$

Then the likelihood function becomes

$$L(\theta|H, T, T, H, H) = \theta^3(1-\theta)^2.$$

Then we choose θ that most likely gives this outcome i.e. maximizes the likelihood. We can find the solution by first taking the logarithm and take the derivative:

$$\begin{aligned}\frac{d}{d\theta}LL(\theta|H, T, T, H, H) &= \frac{d}{d\theta}(3\log\theta + 2\log(1-\theta) \\ &= \frac{3}{\log\theta} - \frac{2}{1-\theta}.\end{aligned}$$

Setting this to zero, we obtain $\hat{\theta} = 3/5$. ◇

In general, the MLE of θ in $\text{Ber}(\theta)$ after observing n independent trials is

$$\frac{n_1}{n},$$

where n_1 is the number of relevant events. For the multinomial distribution $\text{Mul}(\theta_1, \theta_2, \ldots, \theta_k)$ $(\sum_i \theta_i = 1)$ of an event with k possible outcomes $1, 2, \ldots, k$. The MLE based on observing $x_1, x_2, \ldots, x_n$ is

$$\hat{\theta}_i = \frac{n_i}{n_1 + n_2 + \ldots + n_k} \qquad i = 1, 2, \ldots, k.$$

4.2 MLE for Bayesian network

4.2.1 A simple case

Suppose that we have a simple graph $X \to Y$ and data $D = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \{x^0, x^1\}$ and $y_i \in y^0, y^1$. Then the parameters we want to estimate are θ_{x^0}, θ_{x^1}, $\theta_{Y|x^0} = \{\theta_{y^0|x^0}, \theta_{y^1|x^0}\}$ and $\theta_{Y|x^1} = \{\theta_{y^0|x^1}, \theta_{y^1|x^1}\}$. We also denote $\theta_X = \{\theta_{x^0}, \theta_{x^1}\}$, $\theta_{Y|X} = \theta_{Y|x^0} \cup \theta_{Y|x^1}$. and $\theta = \theta_X \cup \theta_{Y|X}$. Assuming that the observations are independent, the likelihood function is

$$\begin{aligned} L(\theta|D) &= \prod_i \mathbb{P}(x_i, y_i|\theta) \\ &= \prod_i \mathbb{P}(x_i|\theta)\mathbb{P}(y_i|x_i, \theta) \\ &= \left(\prod_i \mathbb{P}(x_i|\theta_X)\right)\left(\prod_i \mathbb{P}(y_i|x_i, \theta_{Y|X})\right) \end{aligned}$$

Each individual term can be maximized separately. As in Example 4.1, the MLE of the first term is

$$\hat{\theta}_{x^a} = \frac{n(x^j)}{n(x^0) + n(x^1)} = \frac{n(x^j)}{n},$$

where $n(x^j)$ denotes the number of observations that are equal to x^j. For the second term , we can decompose further.

$$\prod_i \mathbb{P}(y_i|x_i, \theta_{Y|X}) = \prod_{i, x_i = x^0} \mathbb{P}(y_i|x_i, \theta_{Y|x^0}) \prod_{i, x_i = x^1} \mathbb{P}(y_i|x_i, \theta_{Y|x^1})$$

Again, we can maximize each term separately. Let $n(x^0, y^j)$ be the number of (x^0, y^j) in the sample. Then the first term can be computed as

$$\prod_{i, x_i = x^0} \mathbb{P}(y_i|x_i, \theta_{Y|x^0}) = \theta_{y^0|x^0}^{n(x^0, y^0)} \theta_{y^1|x^0}^{n(x^0, y^1)},$$

which, as in Example 4.1, is maximized at

$$\hat{\theta}_{y^j|x^0} = \frac{n(x^0, y^j)}{n(x^0)}.$$

Similarly, we have

$$\hat{\theta}_{y^j|x^1} = \frac{n(x^1, y^j)}{n(x^1)}.$$

4.2.2 General case

Let G be a Bayesian network and $D = \{d_1, d_2, \ldots, d_n\}$ be the data of m variables: $d_i = (x_i^1, x_i^2, \ldots, x_i^m)$. Then by the factorization property of the Bayesian network, the parameters that we need to estimate is

$\theta_{X^j|\mathrm{Pa}_{X^j}}$. Assuming that these parameters for different values of Pa_{X^j} are disjoint, we have

$$\begin{aligned} L(\theta|D) &= \prod_i \mathbb{P}(d_i|\theta) \\ &= \prod_i \prod_j \mathbb{P}(x_i^j|\mathrm{Pa}_{x_i^j}, \theta_{X^j|\mathrm{Pa}_{X^j}}) \\ &= \prod_j \left[\prod_i \mathbb{P}(x_i^j|\mathrm{Pa}_{x_i^j}, \theta_{X_i|\mathrm{Pa}_{X^j}}) \right] \\ &= \prod_j L_j(\theta_{X^j|\mathrm{Pa}_{X^j}}|D) \end{aligned}$$

where L_j is the conditional likelihood of X^j, which can be maximized separately for each j. For example, suppose that all variables have the multinomial conditional distribution. Suppose that a variable X has the set of parents U. Let $\mathrm{Val}(X)$ and $\mathrm{Val}(U)$ denote the set of all possible values of X and U, respectively. Then be can further compute the conditional likelihood:

$$\begin{aligned} L_X(\theta_{X|U}|D) &= \prod_i \theta_{x_i|u_i} \\ &= \prod_{u \in \mathrm{Val}(U)} \left[\prod_{x \in \mathrm{Val}(X)} \theta_{x|u}^{n(x,u)} \right], \end{aligned}$$

where $n(x, u)$ is the number of times that (x, u) appears in D. Again, the maximization can be done for the bracket term for each u, giving the MLE solution as

$$\hat{\theta}_{x|u} = \frac{n(x,u)}{n(u)}.$$

Example 4.2. Let us go back to the Naïve Bayes classifier for spam filtering. Recall that the model is in the form $Y \to X^i$ where X^i runs through all words in the dictionary and $Y \in \{0, 1\}$. From the previous discussion, the MLE based on the model is

$$\hat{\theta}_{X^i|Y} = \frac{n(x^i, y)}{n(y)}$$

Then, to classify an email, we use

$$\begin{aligned} \mathbb{P}(y = 1|x^1, x^2, \ldots, x^m) &= \frac{\mathbb{P}(y=1) \prod_i \mathbb{P}(x^i|y=1)}{\mathbb{P}(y=0) \prod_i \mathbb{P}(x^i|y=0) + \mathbb{P}(y=1) \prod_i \mathbb{P}(x^i|y=1)} \\ &= \frac{\hat{\theta}_{y^1} \prod_i \hat{\theta}_{x^i|y^1}}{\hat{\theta}_{y^0} \prod_i \hat{\theta}_{x^i|y^0} + \hat{\theta}_{y^1} \prod_i \hat{\theta}_{x^i|y^1}}. \end{aligned}$$

◇