Lecture 5: January 10

Lecturer: Donlapark Pornnopparath

5.1 Gaussian Bayesian Networks

Consider a Bayesian networks G and one of the nodes X is continuous with its parent $U = \{U_1, U_2, \ldots, U_k\}$. We can model the conditional distribution of X based on a linear Gaussian:

$$\mathbb{P}(X|u) = N(\beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k; \sigma^2))$$

We want to learn the parameter $\theta_{X|U} = (\beta_0, \beta_1, \ldots, \beta_k, \sigma)$ using MLE. As in the previous lecture, it boils down to computing the likelihood at each node given the data D. The local likelihood function at X is

$$L_X(\theta_{X|U}|D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\beta_0 + \beta_1 u_i^1 + \beta_2 u_i^2 + \ldots + \beta_k u_i^k - x_i)^2\right).$$

Thus the log-likelihood function is

$$LL_X(\theta_{X|U}|D) = \sum_i \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\beta_0 + \beta_1 u_i^1 + \beta_2 u_i^2 + \ldots + \beta_k u_i^k - x_i)^2\right]$$

To find the maximizers, we take the derivative with respect to β_j and set to zero to obtain

$$\overline{xu^j} = \beta_0\overline{u^j} + \beta_1\overline{u^1u^j} + \ldots + \beta_k\overline{u^ku^j}, \qquad j = 0, 1, 2, \ldots, k.$$

where we denoted $u^0 = 1$, $\overline{xu^j} = \frac{1}{n}\sum x_i u_i^j$ and $\overline{u^lu^j} = \frac{1}{n}\sum u_i^l u_i^j$. We can solve the system of $k = 1$ linear equations to get the solution for $\beta_0, \beta_1, \ldots, \beta_k$. The remaining is σ^2, which can be solved by setting $\frac{\partial}{\partial\sigma^2} LL_X(\theta_{X|U}|D) = 0$ and solve for σ^2.

5.2 Bayesian networks with Bayesian parameters

Let us go back to the coin tossing example. Recall that when the coin came up head three out of five times, the MLE would tell us that the parameter is $3/5$. However, we would not be convinced by such conclusion, because we have some *prior* belief about the parameter of the coin even before tossing. And we would be more convinced if instead we tossed the coin $50,000$ times and it turned up head $30,000$. In contrast to the MLE which gives no distinction between $3/5$ and $30,000/50,000$, the *Bayesian statistics* could be used to explain our confidence in the parameter being $\theta = 0.6$ given the latter result compared to the former.

5.2.1 Joint probabilistic model from a given prior

The first component of Bayesian statistics is the *prior distribution* $p(\theta)$ which represents our prior belief about the parameter θ. Suppose that we already have made a choice for $p(\theta)$. Then the joint pdf between

the observed data $x_1, x_2, \ldots, x_n$ and θ is

$$\begin{aligned} p(x_1, \ldots, x_n, \theta) &= p(x_1, \ldots, x_n|\theta)p(\theta) \\ &= p(\theta)\prod_{i=1}^{n} p(x_i|\theta) \\ &= p(\theta)\theta^H(1-\theta)^T, \end{aligned}$$

where H is the number of heads and T is the number of tails. After observing the data, we update our belief about the parameter using the Bayes' rule:

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)} \\ &= \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{\int p(x_1, \ldots, x_n|\theta)p(\theta)\ d\theta}. \end{aligned}$$

This is called the *posterior distribution*. We can use this to compute the chance that next coin toss will turn up head.

$$\begin{aligned} p(x_{n+1}|x_1, \ldots, x_n) &= \int p(x_{n+1}|\theta, x_1, \ldots, x_n)p(\theta|x_1, \ldots, x_n)\ d\theta \\ &= \int p(x_{n+1}|\theta)p(\theta|x_1, \ldots, x_n)\ d\theta. \end{aligned}$$

Assuming that the prior uniformly distributed on $[0, 1]$. Then we can explicitly compute the probability.

$$\begin{aligned} \mathbb{P}(x_{n+1} = 1|x_1, \ldots, x_n) &= \frac{\int_0^1 \theta \cdot \theta^H(1-\theta)^T\ d\theta}{\int_0^1 \theta^H(1-\theta)^T\ d\theta} \\ &= \frac{H+1}{H+T+2}. \end{aligned}$$

Compared to the MLE estimator $H/(H+T)$, the probability are pulled more toward 1 when we get more heads and tails and vice versa.

5.2.2 Choices of prior

One of the most common distribution for prior is the *Beta distribution*

Definition 5.1. A random variable X has a Beta(α_1, α_0) distribution if its pdf is given by

$$p_(\theta) = \gamma\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1},$$

where γ is the normalizing constant: $\gamma = \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}$.

One can interpret the parameters α_0 and α_1 as follows: Consider a single coin toss $X \sim \text{Bern}(\theta)$ where $\theta \sim \text{Beta}(\alpha_1, \alpha_0)$. Then the marginal probability of X is

$$\begin{aligned}\mathbb{P}(X=1) &= \int_0^1 \mathbb{P}(X=1|\theta)p(\theta)\ d\theta \\ &= \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \int_0^1 \theta \cdot \theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\ d\theta \\ &= \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \cdot \frac{\Gamma(\alpha_1+1)\Gamma(\alpha_0)}{\Gamma(\alpha_1+\alpha_0+1)} \\ &= \frac{\alpha_1}{\alpha_1+\alpha_0}.\end{aligned}$$

Therefore, we can view α_1 as the number of imaginary heads and α_0 the number of imaginary tails that we have seen before the experiment. Another important property of the Beta prior is that the corresponding posterior is also a Beta distribution.

$$\begin{aligned}p(\theta|x_1,\dots,x_n) &\propto \mathbb{P}(x_1,\dots,x_n|\theta)p(\theta) \\ &\propto \theta^H(1-\theta)^T \cdot \theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} \\ &= \theta^{\alpha_1+H-1}(1-\theta)^{\alpha_0+T-1},\end{aligned}$$

which is a pdf of $\text{Beta}(\alpha_1+H, \alpha_0+T)$. This leads to

$$\mathbb{P}(X_{n+1}=1|x_1,\dots,x_n) = \frac{\alpha_1+H}{\alpha_1+\alpha_0+H+T},$$

which tells us that we have seen α_1+H real and imaginary heads and α_0+T real and imaginary tails. In general, when $X_1, X_2, \dots, X_k \sim \text{Bin}(n,\theta)$, the posterior distribution is $\text{Beta}\left(\alpha_1+\sum_i x_i, \alpha_0+\sum_i kn-\sum_i x_i\right)$. In Bayesian analysis, we are always interested in an experiment setting where the prior and posterior distribution belong to the same family.

Definition 5.2. From observed data D, we have the following relation between prior distribution, likelihood function and the posterior distribution:

$$p(\theta|D) = L(\theta|D)p(\theta).$$

If the probability distribution functions $p(\theta)$ and $p(\theta|D)$ belong to the same family, then they are called *conjugate distributions* and the prior distribution is called the *conjugate prior*.

Thus, from the previous example, the Beta distribution is a conjugate prior to the binomial likelihood. Now we consider a more general situation where there are m independent trials, each of which has one of d outcomes $1, 2, \dots, d$. For an observed data $\boldsymbol{x} = (n_1, n_2, \dots, n_d)$ where $\sum_j n_j = m$, we have the multinomial likelihood:

$$L(\boldsymbol{\theta}|\boldsymbol{x}) \propto \boldsymbol{\theta}^{\boldsymbol{x}} = \theta_1^{n_1}\theta_2^{n_2}\dots\theta_d^{n_d},$$

which has the conjugate prior as the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \ \text{ if } p(\theta) \propto \prod_{i=1}^d \theta_i^{\alpha-1}.$$

After observing data $D = \{x_1, \dots, x_n\}$, it has the posterior distribution

$$\boldsymbol{\theta}_D \sim \text{Dirichlet}\left(\boldsymbol{\alpha} + \sum_{i=1}^n x_i\right)$$

Similarly to the Beta prior, the probability of the next outcome has a simple interpretation:

$$\mathbb{P}(x_{n+1} = j|D) = \frac{n(j) + \alpha_j}{mn + \alpha},$$

where $n(j)$ is the number of j's in the sample. As before, the posterior tells us about the proportion of real and imaginary j's that came up to the total number of outcomes in the experiment.