Lecture 23: November 5

Lecturer: Donlapark Pornnopparath

23.1 Bayesian inference

In contrast to the previous approach where we took parameter θ as a fixed unknown quantity, the Bayesian approach treats it as a random variable. The distribution of θ is shaped by our beliefs prior to experiments and represented in the form of pdf $\pi(\theta)$. After observing data $x_1, x_2, \ldots, x_n$, the distribution of θ is updated with this new information into $\pi(\theta|x_1, x_2, \ldots, x_n)$ using Bayes' rule. There are three components of this updating process:

- **Prior distribution**: $\pi(\theta)$
- **Likelihood function**: $f(\boldsymbol{x}|\theta)$
- **Posterior distribution**: $\pi(\theta|\boldsymbol{x})$

By the Bayes' rule,

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{q(x)},$$

where $q(x)$ is the marginal density of $\boldsymbol{X}$:

$$q(x) = \int f(\boldsymbol{x}|\theta)\pi(\theta)\ d\theta. \tag{23.1}$$

Example 23.1. Suppose you toss a coin 10 times and it turns up head every time, then you might think that the coin is weighted. This can be formulated in Bayesian language as follows: For the number of heads $x = 10$ that we observed, we have

- **Prior distribution**: $\pi(p) = \text{unif}[0, 1]$
- **Likelihood function**: $f(x|p) = \binom{10}{x}p^x(1-p)^{10-x} = p^{10}$.

The marginal distribution is

$$q(10) = \int_0^1 p^{10}\ dp = \frac{p^{11}}{11}\Big|_0^1 = \frac{1}{11}.$$

Hence, the posterior distribution is

$$\pi(p|10) = \frac{f(10|p)\pi(p)}{q(10)} = \frac{p^{10}}{1/11} = 11p^{10}.$$

We can then use this to calculate the chance of coin being unfair.

$$\mathbb{P}\left(p > \frac{1}{2}\Big|x = 10\right) = \int_{1/2}^1 11p^{10}\ dp = p^{11}\Big|_{1/2}^1 = 1 - \frac{1}{2^{11}} \approx 0.9995.$$

$\diamond$

Example 23.2. We consider the previous example with general n, p and number of heads x. Thus, $X \sim \text{Bin}(n, p)$. Again, assume that p follows the distribution unif$[0, 1]$. Then

$$\pi(p|x) = \frac{p^x(1-p)^{n-x}}{\int_0^1 p^x(1-p)^{n-x}\ dx}.$$

This is related to the beta distribution Beta(a, b), whose pdf is defined by

$$\lambda(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1},$$

where the normalizing factor is the *Beta function*

$$B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}\ dp = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Using the above notations, the posterior distribution is Beta$(x+1, n-x+1)$. Note that the prior distribution unif$[0, 1]$ can be written as Beta$(1, 1)$. When the prior and the posterior distribution belong to the same family, we say that they are *conjugate distributions* and the prior is called a *conjugate prior* to the likelihood function, which is the binomial distribution in this case.

If instead we have Beta(a, b) as a prior distribution for p, then

$$\begin{aligned}\pi(p|x) &\propto f(x|p)\pi(p) \\ &\propto \left[p^x(1-p)^{n-x}\right]\left[p^{a-1}(1-p)^{b-1}\right] \\ &= p^{a+x-1}(1-p)^{b+n-x-1},\end{aligned}$$

which is a pdf of a Beta$(a+x, b+n-x)$ distribution. We can see that the beta distributions can be updated using a simple rule of increasing a by the number of successes and b by the number of failures. ◇

Example 23.3. Let $X \sim N(\mu, \sigma^2)$, where σ^2 is fixed and the prior distribution of μ is $N(\theta_0, \tau_0^2)$ Then it can be showed that the posterior distribution is also normal: $\pi(\mu|x) = N(\theta_1, \tau_1^2)$ where the parameters is given by

$$\theta_1 = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} x + \frac{\sigma^2}{\tau_0^2 + \sigma^2}\theta_0 \tag{23.2}$$

$$\tau_1^2 = \frac{\tau_0^2\sigma^2}{\tau_0^2 + \sigma^2}. \tag{23.3}$$

Thus, the updated mean is the weighted average between the prior mean θ_0 and the sample x, and the distances to these values are proportional to their respective variances. ◇

In the classical approach, we tried to find a good estimator that minimizes a risk function. The Bayesian approach also relies on this concept.

Definition 23.4. Let θ be a parameter with prior distribution $\pi(\theta)$, δ be an estimator of θ and $R(\theta, \delta)$ be a risk function of δ. The *Bayes risk* of δ is defined by

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\ d\theta. \tag{23.4}$$

An estimator that minimizes (23.4) is called a *Bayes estimator*.

The task of finding a minimizer can be simplified by considering the Bayes risk: Assuming that the risk function is defined based on the loss function $L(\theta, \delta(x))$

$$\begin{aligned}\int R(\theta,\delta)\pi(\theta)\ d\theta &= \int\int L(\theta,\delta(x))f(x|\theta)\pi(\theta)\ dxd\theta \\ &= \int\int L(\theta,\delta(x))\pi(\theta|x)g(x)\ dxd\theta \\ &= \int \mathbb{E}[L(\theta,\delta(X)|X]g(x)\ dx,\end{aligned}$$

where $g(x)$ is given in (23.1). If we can find a minimizer $\delta_\pi(x)$ of the *posterior risk* $\mathbb{E}[L(\theta, \delta(X)|X = x]$ for each x, then $\delta_\pi(X)$ is a Bayes estimator since

$$\int R(\theta,\delta)\pi(\theta)\ d\theta = \int \mathbb{E}[L(\theta,\delta(X)|X]g(x)\ dx \geq \int \mathbb{E}[L(\theta,\delta_\pi(X)|X]g(x)\ dx = \int R(\theta,\delta_\pi)\pi(\theta)\ d\theta.$$

We give an example to illustrate how we can utilize this method.

Example 23.5. Consider the squared error loss $L(\theta, \delta) = (\delta - g(\theta))^2$. From the previous discussion, we want to minimize

$$\begin{aligned}\mathbb{E}[(\delta - g(\theta))^2|X = x] &= \delta^2\mathbb{E}[1|X = x] - 2\delta\mathbb{E}[g(\theta)|X = x] + \mathbb{E}[(g(\theta))^2|X = x] \\ &= \delta^2 - 2\delta\mathbb{E}[g(\theta)|X = x] + \mathbb{E}[(g(\theta))^2|X = x].\end{aligned}$$

Taking the derivative with respect to d and setting equal to zero, we see that

$$\delta_\pi(x) = \mathbb{E}[g(\theta)|X = x].$$

is a Bayes estimator. For a concrete example, we take Example 23.3 with $\theta = \mu$ and $g(\mu) = \mu^2$. The minimizer is

$$\mathbb{E}[\mu^2|\theta_1, \tau_1^2] = \mathrm{Var}(\mu|\theta_1, \tau_1^2) + \left(\mathbb{E}[\mu|\theta_1, \tau_1^2]\right)^2 = \theta_1^2 + \tau_1^2,$$

where the posterior parameter θ_1 and τ_1^2 are given in (23.2) and (23.3), repectively. Therefore, the Bayes estimator is

$$\delta(X) = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2}X + \frac{\sigma^2}{\tau_0^2 + \sigma^2}\theta_0\right)^2 + \frac{\tau_0^2\sigma^2}{\tau_0^2 + \sigma^2}.$$

Going back to the binomial examples Example 23.2, in order to estimate $g(p) = p$, we have to use the fact that for any $X \sim \mathrm{Beta}(a, b)$,

$$\mathbb{E}X = \frac{a}{a+b}.$$

Since the posterior beta distribution is $\mathrm{Beta}(a + x, b + n - x)$, this means that the Bayes estimator is given by

$$\mathbb{E}[p|X] = \frac{X + a}{n + a + b} = \left[\frac{n}{n + a + b}\right]\frac{X}{n} + \left[1 - \frac{n}{n + a + b}\right]\frac{a}{a + b},$$

which is a weighted average between the prior mean $a/(a + b)$ and the UMVU estimator X/n of p. ◇

Example 23.6. Consider the zero-one loss function:

$$L(\theta,\delta) = \begin{cases} 0 & \theta = \delta \\ 1 & \text{otherwise} \end{cases}.$$

Then

$$\mathbb{E}[L(\theta,\delta)|X = x] = \sum_{\theta\neq\delta}\pi(\theta|x) = 1 - \pi(\delta|x),$$

which can be minimized by choosing $\delta(x)$ that maximizes $\pi(\delta(x)|x)$ i.e. δ is the maximum a posteriori estimator (MAP) of θ. ◇