Lecture 8

Bayesian Inference.

1 Frequentists and Bayesian Paradigms

According to the frequentists theory, it is assumed that unknown parameter θ is some fixed number or vector. Given parameter value θ, we observe sample data X from distribution $f(\cdot|\theta)$. To estimate parameter θ, we introduce an estimator $T(X)$ which is a statistic, i.e. function of the data. This statistic is a random variable, since X. That is, the randomness here comes from randomness of sampling. Different properties of estimation characterize this randomness. For example, the concept of unbiasness means that if we observe many samples from distribution $f(\cdot|\theta)$, then the average of $T(X)$ over the samples will be close to the true parameter value θ, i.e. $E_\theta[T(X)] = \theta$. The concept of consistency means that if we observe many samples from distribution $f(\cdot|\theta)$, then the distribution of $T(X)$ over these samples will be close in probability to the true parameter value θ, at least when we observe large samples, i.e. $T(X)$ will be in a small neighborhood of θ in most observed samples.

In contrast, Bayesian theory assumes that θ is some random variable and we are interested in the realization of this random variable. This realization, say, θ_0, is thought to be the true parameter value. It is assumed that we know the distribution of θ, or at least its approximation. This distribution is called a prior. It usually comes from our subjective belief based on our past experience. Once θ_0 is realized, we observe a random sample $X = (X_1, ..., X_n)$ from distribution $f(\cdot|\theta_0)$. Once we have data, the best thing we can do is to calculate conditional distribution of θ given $X_1, ..., X_n$. This conditional distribution is called the *posterior*. The posterior is used to create an estimate for θ. Since we condition on the observations X_i, we treat them as given. The randomness in posterior is an uncertainty about θ. The Bayesian approach is often used in the learning theory.

2 Bayesian Updating

Let $\pi(\theta)$ denote our prior. In other words, parameter of interest θ is a random variable with distribution $\pi(\cdot)$. Our model is $f(\cdot|\theta)$. In other words, once we have a realization of a parameter value θ, the observed sample data X are drawn from the conditional distribution $f(\cdot|\theta)$. Then the joint pdf of θ and X is $f(x, \theta) = \pi(\theta)f(x|\theta)$. The prior predictive distribution is $m(x) = \int \pi(\tilde{\theta})f(x|\tilde{\theta})d\tilde{\theta}$. Thus, the prior predictive distribution means the marginal pdf of our sample data X. Once we observe $X = x$, we can calculate the

posterior distribution for θ as

$$\pi(\theta|X=x) = \frac{f(x,\theta)}{m(x)} = \frac{\pi(\theta)f(x|\theta)}{\int \pi(\tilde{\theta})f(x|\tilde{\theta})d\tilde{\theta}}.$$

Example Let $X = (X_1, ..., X_n)$ be a random sample from a Bernoulli(p) distribution. Then the joint pdf of the data is $f(x|p) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$ where $x = (x_1, ..., x_n)$. In classical theory, we would consider some estimator of p. For example, we can take $T(X) = \overline{X}_n$. Then we have already seen that $E_p[T(X)] = p$ and $T(X) \to_p p$. In Bayesian theory we need some prior distribution of p. For example, suppose we believe that all p are equally possible. Then we have a uniform prior, i.e. $\pi(p) = 1$ if $p \in [0,1]$ and 0 otherwise. Then the joint pdf of p and X is $f(x,p) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} I(0 \le p \le 1)$. The prior predictive distribution is

$$m(x) = \int_0^1 p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} dp = B(\sum_{i=1}^n x_i + 1, n - \sum_{i=1}^n x_i + 1),$$

where $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ is the Beta-function. The posterior distribution is

$$\pi(p|X=x) = \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}{B(\sum_{i=1}^n x_i + 1, n - \sum_{i=1}^n x_i + 1)} I(0 \le p \le 1)$$

This distribution is called *Beta* $\mathcal{B}(\alpha, \beta)$ distribution with parameters $\alpha = \sum_{i=1}^n x_i$ and $\beta = n - \sum_{i=1}^n x_i + 1$. Its mean is

$$E[p|X=x] = \frac{\alpha}{(\alpha+\beta)} = \frac{\sum_{i=1}^n x_i + 1}{n+2}.$$

Its variance is

$$V(p|X=x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{(\sum_{i=1}^n x_i + 1)(n - \sum_{i=1}^n x_i + 1)}{(n+2)^2(n+3)}.$$

2.1 How to calculate posterior distribution

Here we consider a trick that may help to calculate the posterior analytically. It does not work always, though.

Let $X = (X_1, ...X_n)$ be a random sample from an $N(\mu, \sigma^2)$ distribution. Suppose σ^2 is known. Let $N(\mu_0, \tau^2)$ be the prior distribution for μ. Then the posterior distribution is $\pi(\mu|X=x) = \pi(\mu)f(x|\mu)/m(x)$ where $\pi(\mu)$ denotes the prior distribution and $m(x)$ denotes the prior predictive distribution. Note that $m(x)$ does not depend on μ. So $m(x)$ is just a constant that normalizes $\int \pi(\mu|X=x)d\mu$ to 1. Therefore, when we calculate $\pi(\mu)f(x|\mu)$, we can denote all multiplicative terms which do not contain μ as some constant C instead of keeping track of all these terms. Once we have an expression for $\pi(\mu)f(x|\mu)$ as a function of μ, we can integrate it in order to find the normalizing constant $m(x)$. In our case,

$$\pi(\mu|X=x) = C\pi(\mu)f(x|\mu) = C\exp\{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2 - (\mu-\mu_0)^2/(2\tau^2)\}$$

where C contains all terms which do not contain μ. Thus,

$$\begin{aligned}
\pi(\mu|X=x) &= C\exp\left\{\frac{\mu\sum_{i=1}^n x_i}{\sigma^2}-\frac{n\mu^2}{2\sigma^2}-\frac{\mu^2}{2\tau^2}+\frac{\mu\mu_0}{\tau^2}\right\}\\
&= C\exp\left\{-\mu^2\left(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2}\right)+2\mu\left(\frac{\sum_{i=1}^n x_i}{2\sigma^2}+\frac{\mu_0}{2\tau^2}\right)\right\}\\
&= C\exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\mu^2-2\mu\left(\frac{\sum_{i=1}^n x_i/\sigma^2+\mu_0/\tau^2}{n/\sigma^2+1/\tau^2}\right)\right]\right\}\\
&= C\exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)[\mu-\tilde{\mu}]^2\right\}\\
&= C\exp\left\{-\frac{(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right\},
\end{aligned}$$

where $\tilde{\mu}=(\sum_{i=1}^n x_i/\sigma^2+\mu_0/\tau^2)/(n/\sigma^2+1/\tau^2)$ and $\tilde{\sigma}^2=1/(n/\sigma^2+1/\tau^2)$. Note that, via some abuse of notation, one letter C actually stands for different constants. Thus, the conditional distribution of μ given $X=x$ is $N(\tilde{\mu},\tilde{\sigma}^2)$. Now, it is easy to find a constant in the last expression, namely $C=1/(\sqrt{2\pi}\tilde{\sigma})$.

Note that in the example above, posterior mean $\tilde{\mu}$ is a weighted average of sample average $\overline{x}_n$ and prior mean μ_0, i.e. $\tilde{\mu}=\omega_1\overline{x}_n+\omega_2\mu_0$ with $\omega_1+\omega_2=1$ where $\omega_1=(n/\sigma^2)/(n/\sigma^2+1/\tau^2)$ and $\omega_2=(1/\tau^2)/(n/\sigma^2+1/\tau^2)$. Here $1/\tau^2$ may be interpreted as the precision of initial information. If initial information is very precise, i.e. $1/\tau^2$ is large (or τ^2 is small), then the prior mean gets almost all weight, and $\tilde{\mu}$ is close to μ_0. Thus, we almost ignore new information in the form of a sample $X_1,...,X_n$. If initial information is poor, i.e. $1/\tau^2$ is small, prior mean gets almost no weight, and $\tilde{\mu}$ is close to $\overline{x}_n$. Note that as $n\to\infty$, information from the sample dominates prior information and $\tilde{\mu}\to\overline{x}_n$. Moreover, at least in our example, as $n\to\infty$, $\tilde{\sigma}^2\to 0$. Thus, as the sample size increases, the posterior distribution converges to a degenerate distribution concentrated at the true parameter value.

Once we have the posterior distribution, we can construct an estimator of the parameter of interest. Common examples include posterior mean and posterior mode (posterior mode denotes the point with the greatest pdf value on the posterior distribution).

An important problem with Bayesian estimation is that if a prior distribution puts zero probability mass on the true parameter value, then no matter how large our sample is, posterior distribution will put zero mass on the true parameter value as well. You will see examples of this phenomenon in 14.384.

2.2 Conjugate Classes

Let $\mathcal{F}$ be the class of distributions indexed by θ. Let $\mathcal{P}$ be the class of prior distributions of θ. Then we say that $\mathcal{P}$ is conjugate to $\mathcal{F}$ if whenever data is distributed according to $\mathcal{F}$ and prior distribution is from $\mathcal{P}$, then the posterior distribution is from $\mathcal{P}$ as well. For example, we have already seen that the class of normal distributions is conjugate to the class of normal distributions with known (fixed) variance. It is also known that the class of $\mathcal{B}$-distributions is conjugate to the class of binomial distributions. The concept of conjugate classes is introduced because of its mathematical convenience. It is relatively easy to calculate the posterior when the prior lies in the conjugate class. Conjugate priors were almost the only priors used for a long time, as the others tend to be not analytically tractable. Nowadays, there are numerical algorithms

(MCMC- Markov Chain Monte-Carlo), that allow one to calculate posterior for priors outside the conjugate family. MCMC will be discussed in 14.384 and 14.385.

2.3 Credible Intervals

The posterior distribution contains all the information that a researcher can get from the data. However, it is often impractical to report a posterior distribution, as it might be intractable or it might not have analytic form at all. Therefore, it is a common practice to report only some characteristics of a posterior distribution, such as its mean, variance, and mode. Mean and mode of the posterior serve as point estimators of the parameter while variance shows the precision of the posterior distribution. A better way to show the precision of the posterior distribution is the concept of *credible intervals*. Let X be our data, $\theta \in \Theta$ our parameter, and $\pi(\theta|X = x)$ the posterior distribution. Then for any $\gamma \in [0, 1]$, a set $C(x) \subset \Theta$ is called γ-credible if $\pi(\theta \in C(x)|X = x) \geq \gamma$. In words, set $C(x)$ contains true parameter value θ with probability of at least γ. Of course, the whole parameter space Θ is γ-credible. But, apparently, reporting Θ as a γ-credible set is not useful at all. Therefore we should try to choose the smallest γ-credible set. The smallest γ-credible set contains only points with the highest posterior density.

As an example, let $X = (X_1, ..., X_n)$ be a random sample from $N(\mu, \sigma^2)$ with σ^2 known and let $N(\mu_0, \tau^2)$ be the prior distribution of μ. We have already seen that the conditional distribution of μ given $X = x$ is $N(\tilde{\mu}, \tilde{\sigma}^2)$ with $\tilde{\mu}$ and $\tilde{\sigma}^2$ defined as above. Then $(\mu - \tilde{\mu})/\tilde{\sigma} \sim N(0, 1)$. Let $z_{(1+\gamma)/2}$ be the $(1 + \gamma)/2$-quantile of the standard normal distribution. Then $\pi\{(\mu - \tilde{\mu})/\tilde{\sigma} \in [-z_{(1+\gamma)/2}, z_{(1+\gamma)/2}]|X = x\} = \gamma$ or, equivalently, $\pi\{\tilde{\mu} - z_{(1+\gamma)/2}\tilde{\sigma} \leq \mu \leq \tilde{\mu} + z_{(1+\gamma)/2}\tilde{\sigma}|X = x\} = \gamma$. Since the pdf $\phi(x)$ of the standard normal distribution is decreasing on $x \geq 0$ and increasing on $x \leq 0$, $[\tilde{\mu} - z_{(1+\gamma)/2}\tilde{\sigma}, \tilde{\mu} + z_{(1+\gamma)/2}\tilde{\sigma}]$ is the shortest γ-credible interval.

2.4 Bayesian Testing

Let $\pi(\theta|X = x)$ be our posterior distribution. How can we test the hypothesis that $\theta = \theta_0$ against the hypothesis that $\theta \neq \theta_0$? One idea is to look at $\pi(\theta_0|X = x)$. If it is small then we conclude that there is evidence against the hypothesis that $\theta = \theta_0$. But what does "small" mean? Let $\hat{\Theta}$ be the set of all θ such that $\pi(\theta|X = x) \leq \pi(\theta_0|X = x)$. Then the posterior probability that $\theta \in \hat{\Theta}$, i.e. $\pi(\theta \in \hat{\Theta}|X = x)$, is called a Bayesian p-value. If a Bayesian p-value is small, then it is unlikely that θ_0 is the true parameter value. So we conclude in this case that we reject the hypothesis that $\theta = \theta_0$ in favor of the hypothesis that $\theta \neq \theta_0$. It is a common practice to compare Bayesian p-value with 0.01, 0.05 or 0.10, i.e. we reject the hypothesis that $\theta = \theta_0$ if the Bayesian p-value is smaller than 0.01, 0.05 or 0.10.

As an example, let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$ with σ^2 known, and let $N(\mu_0, \tau^2)$ be the prior distribution of μ. Then the conditional distribution of μ given $X = x$ is $N(\tilde{\mu}, \tilde{\sigma}^2)$ with $\tilde{\mu}$ and $\tilde{\sigma}^2$ defined as above. Suppose we want to test the hypothesis that $\mu = 5$. Then Bayesian p-value is $2(1 - \Phi((5 - \tilde{\mu})/\tilde{\sigma}))$.

14.381 Statistical Method in Economics
Fall 2013