STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 20 — Bayesian analysis

Our treatment of parameter estimation thus far has assumed that θ is an unknown but non-random quantity—it is some fixed parameter describing the true distribution of data, and our goal was to determine this parameter. This is the called the **frequentist** paradigm of statistical inference. In this and the next lecture, we will describe an alternative **Bayesian** paradigm, in which θ itself is modeled as a random variable. The Bayesian paradigm naturally incorporates our prior belief about the unknown parameter θ, and updates this belief based on observed data.

20.1 Prior and posterior distributions

Recall that if X, Y are two random variables having joint PDF or PMF $f_{X,Y}(x,y)$, then the **marginal distribution** of X is given by the PDF

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

in the continuous case and by the PMF

$$f_X(x) = \sum_y f_{X,Y}(x,y)$$

in the discrete case; this describes the probability distribution of X alone. The **conditional distribution** of Y given $X = x$ is defined by the PDF or PMF

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

and represents the probability distribution of Y if it is known that $X = x$. (This is a PDF or PMF as a function of y, for any fixed x.) Defining similarly the marginal distribution $f_Y(y)$ of Y and the conditional distribution $f_{X|Y}(x|y)$ of X given $Y = y$, the joint PDF $f_{X,Y}(x,y)$ factors in two ways as

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y).$$

In Bayesian analysis, before data is observed, the unknown parameter is modeled as a random variable Θ having a probability distribution $f_\Theta(\theta)$, called the **prior distribution**. This distribution represents our prior belief about the value of this parameter. Conditional on $\Theta = \theta$, the observed data X is assumed to have distribution $f_{X|\Theta}(x|\theta)$, where $f_{X|\Theta}(x|\theta)$ defines a parametric model with parameter θ, as in our previous lectures.[1] The joint distribution of Θ and X is then the product

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_\Theta(\theta),$$

[1] For notational simplicity, we are considering here a single data value X, but this extends naturally to the case where $\mathbf{X} = (X_1, \ldots, X_n)$ is a data vector and $f_{X|\Theta}(\mathbf{x}|\theta)$ is the joint distribution of $\mathbf{X}$ given θ.

and the marginal distribution of X (in the continuous case) is

$$f_X(x) = \int f_{X,\Theta}(x,\theta)d\theta = \int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta.$$

The conditional distribution of Θ given $X = x$ is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta')f_\Theta(\theta')d\theta'}. \tag{20.1}$$

This is called the **posterior distribution** of Θ: It represents our knowledge about the parameter Θ after having observed the data X. We often summarize the preceding equation simply as

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)f_\Theta(\theta) \tag{20.2}$$

$$\text{Posterior density} \propto \text{Likelihood} \times \text{Prior density}$$

where the symbol $\propto$ hides the proportionality factor $f_X(x) = \int f_{X|\Theta}(x|\theta')f_\Theta(\theta')d\theta'$ which does not depend on θ.

Example 20.1. Let $P \in (0,1)$ be the probability of heads for a biased coin, and let $X_1, \ldots, X_n$ be the outcomes of n tosses of this coin. If we do not have any prior information about P, we might choose for its prior distribution Uniform$(0,1)$, having PDF $f_P(p) = 1$ for all $p \in (0,1)$. Given $P = p$, we model $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p). Then the joint distribution of $P, X_1, \ldots, X_n$ is given by

$$\begin{aligned} f_{X,P}(x_1,\ldots,x_n,p) &= f_{X|P}(x_1,\ldots,x_n|p)f_P(p) \\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \times 1 = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

Let $s = x_1 + \ldots + x_n$. The marginal distribution of $X_1, \ldots, X_n$ is obtained by integrating $f_{X,P}(x_1,\ldots,x_n,p)$ over p:

$$f_X(x_1,\ldots,x_n) = \int_0^1 p^s(1-p)^{n-s}dp = B(s+1, n-s+1)$$

where $B(x,y)$ is the Beta function

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Hence the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ has PDF

$$f_{P|X}(p|x_1,\ldots,x_n) = \frac{f_{X,P}(x_1,\ldots,x_n,p)}{f_X(x_1,\ldots,x_n)} = \frac{1}{B(s+1,n-s+1)}p^s(1-p)^{n-s}.$$

This is the PDF of the Beta$(s+1, n-s+1)$ distribution[2], so the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ is Beta$(s+1, n-s+1)$, where $s = x_1 + \ldots + x_n$.

[2] The Beta(α,β) distribution is a continuous distribution on $(0,1)$ with PDF $f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$.

We computed explicitly the marginal distribution $f_X(x_1, \ldots, x_n)$ above, but this was not necessary to arrive at the answer. Indeed, equation (20.2) gives

$$f_{P|X}(p|x_1, \ldots, x_n) \propto f_{X|P}(x_1, \ldots, x_n|p) f_P(p) = p^s(1-p)^{n-s}.$$

This tells us that the PDF of the posterior distribution of P is proportional to $p^s(1-p)^{n-s}$, as a function of p. Then it must be the PDF of the Beta$(s+1, n-s+1)$ distribution, and the proportionality constant must be whatever constant is required to make this PDF integrate to 1 over $p \in (0,1)$. We will repeatedly use this trick to simplify our calculations of posterior distributions.

Example 20.2. Suppose now we have a prior belief that P is close to $1/2$. There are various prior distributions that we can choose to encode this belief; it will turn out to be mathematically convenient to use the prior distribution Beta(α, α), which has mean $1/2$ and variance $1/(8\alpha+4)$. The constant α may be chosen depending on how confident we are, a priori, that P is near $1/2$—choosing $\alpha = 1$ reduces to the Uniform$(0,1)$ prior of the previous example, whereas choosing $\alpha > 1$ yields a prior distribution more concentrated around $1/2$.

The prior distribution Beta(α, α) has PDF $f_P(p) = \frac{1}{B(\alpha,\alpha)} p^{\alpha-1}(1-p)^{\alpha-1}$. Then, applying equation (20.2), the posterior distribution of P given $X_1 = x_1, \ldots, X_n = x_n$ has PDF

$$\begin{aligned} f_{P|X}(p|x_1, \ldots, x_n) &\propto f_{X|P}(x_1, \ldots, x_n|p) f_P(p) \\ &\propto p^s(1-p)^{n-s} \times p^{\alpha-1}(1-p)^{\alpha-1} = p^{s+\alpha-1}(1-p)^{n-s+\alpha-1}, \end{aligned}$$

where $s = x_1 + \ldots + x_n$ as before, and where the symbol $\propto$ hides any proportionality constants that do not depend on p. This is proportional to the PDF of the distribution Beta$(s+\alpha, n-s+\alpha)$, so this Beta distribution is the posterior distribution of P.

In the previous example, the parametric form for the prior was (cleverly) chosen so that the posterior would be of the same form—they were both Beta distributions. This type of prior is called a **conjugate prior** for P in the Bernoulli model. Use of a conjugate prior is mostly for mathematical and computational convenience—in principle, any prior $f_P(p)$ on $(0,1)$ may be used. The resulting posterior distribution may be not be a simple named distribution with a closed-form PDF, but the PDF may be computed numerically from equation (20.1) by numerically evaluating the integral in the denominator of this equation.

Example 20.3. Let $\Lambda \in (0, \infty)$ be the parameter of the Poisson model $X_1, \ldots, X_n \overset{IID}{\sim}$ Poisson(λ). As a prior distribution for Λ, let us take the Gamma distribution Gamma(α, β). The prior and likelihood are given by

$$\begin{aligned} f_\Lambda(\lambda) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\ f_{X|\Lambda}(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}. \end{aligned}$$

Dropping proportionality constants that do not depend on λ, the posterior distribution of Λ given $X_1 = x_1, \ldots, X_n = x_n$ is then

$$f_{\Lambda|X}(\lambda|x_1, \ldots, x_n) \propto f_{X|\Lambda}(x_1, \ldots, x_n|\lambda) f_\Lambda(\lambda) \propto \prod_{i=1}^n (\lambda^{x_i} e^{-\lambda}) \times \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{s+\alpha-1} e^{-(n+\beta)\lambda},$$

where $s = x_1 + \ldots + x_n$. This is proportional to the PDF of the Gamma$(s+\alpha, n+\beta)$ distribution, so the posterior distribution of Λ must be Gamma$(s+\alpha, n+\beta)$.

As the prior and posterior are both Gamma distributions, the Gamma distribution is a conjugate prior for Λ in the Poisson model.

20.2 Point estimates and credible intervals

To the Bayesian statistician, the posterior distribution is the complete answer to the question: What is the value of θ? In many applications, though, we would still like to have a single estimate $\hat{\theta}$, as well as an interval describing our uncertainty about θ.

The **posterior mean** and **posterior mode** are the mean and mode of the posterior distribution of Θ; both of these are commonly used as a Bayesian estimate $\hat{\theta}$ for θ. A $100(1-\alpha)\%$ **Bayesian credible interval** is an interval I such that the posterior probability $\mathbb{P}[\Theta \in I \mid X] = 1-\alpha$, and is the Bayesian analogue to a frequentist confidence interval. One common choice for I is simply the interval $[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}]$ where $\theta^{(\alpha/2)}$ and $\theta^{(1-\alpha/2)}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution of Θ. Note that the interpretation of a Bayesian credible interval is different from the interpretation of a frequentist confidence interval—in the Bayesian framework, the parameter Θ is modeled as random, and $1-\alpha$ is the probability that this random parameter Θ belongs to an interval that is fixed conditional on the observed data.

Example 20.4. From Example 20.2, the posterior distribution of P is Beta$(s+\alpha, n-s+\alpha)$. The posterior mean is then $(s+\alpha)/(n+2\alpha)$, and the posterior mode is $(s+\alpha-1)/(n+2\alpha-2)$. Both of these may be taken as a point estimate $\hat{p}$ for p. The interval from the 0.05 to the 0.95 quantile of the Beta$(s+\alpha, n-s+\alpha)$ distribution forms a 90% Bayesian credible interval for p.

Example 20.5. From Example 20.3, the posterior distribution of Λ is Gamma$(s+\alpha, n+\beta)$. The posterior mean and mode are then $(s+\alpha)/(n+\beta)$ and $(s+\alpha-1)/(n+\beta)$, and either may be used as a point estimate $\hat{\lambda}$ for λ. The interval from the 0.05 to the 0.95 quantile of the Gamma$(s+\alpha, n+\beta)$ distribution forms a 90% Bayesian credible interval for λ.