Note on Hierarchical Models

1 Introduction

Let P be a probability on some sample space S, and let A and B be two events in S (i.e., $A \subset S$ and $B \subset S$). The conditional probability of A given B is denoted by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1}$$

and by the same reasoning, we have the conditional probability of B given A:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}. \tag{2}$$

Combining (1) and (2), we obtain

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}. \tag{3}$$

Heuristically, the equation (3) says that, we can compute the probability of A given B, if we have the conditional probability of B given A along with the probabilities of both A and B.

2 Bayes Theorem

Now, we ask if we can obtain the conditional probability $P(A|B)$ in (3) without the knowledge of $P(B)$ on the denominator. The answer is yes (but you need to know some other quantities), and the key here is a clever manipulation of conditional probability.

Let A^c denote a complement of A. It is easily seen that

$$P(B) = P((B \cap A) \cup (B \cap A^c)) = P(B \cap A) + P(B \cap A^c)$$

(use the Venn diagram), and by using (2), we can extend the above by

$$P(B) = P(B \cap A) + P(B \cap A^c) = P(B|A)P(A) + P(B|A^c)P(A^c) \tag{4}$$

The equation (4) is called the law of total probability.

Then, combining (3) and (4), we obtain

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{5}$$

so that you do not need to know $P(B)$, but then you have to know the value of $P(B|A^c)$ instead, for (5) to work.

The equation (5) is called the Bayes theorem, in its simplest form.

In general, the $P(B)$ on the denominator of (5) can be expressed as an arbitrary summation and even as an integration. Note that the above equation $P(B) = P(B \cap A) + P(B \cap A^c)$ worked because $A \cup A^c = S$, and A and A^c are disjoint. So if we divide S into disjoint sets $A_1, \ldots A_n$ such that $A_1 \cup \cdots \cup A_n = S$, then we have, as an analogue of (4),

$$\begin{aligned} P(B) &= P(B \cap A_1) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n) \end{aligned}$$

and then we have a more general Bayes theorem

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)} \\ &= \frac{P(B|A)P(A)}{\sum_{i=1}^n P(B|A_i)P(A_i)}. \end{aligned} \tag{6}$$

Again, the only requirement for (6) to work is that $A_1 \cup \cdots \cup A_n = S$ and that the sets A_i are disjoint.

Note that we can write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A) \tag{7}$$

because the quantity A in $P(A|B)$ is of central interest to us, and the denominator $P(B)$ does not depend on A.

Here, we have $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$. So we can think of it as A term is being "integrated out." This concept will become clearer in the next section.

3 Bayes Theorem for Probability Densities

Suppose now that there are two random variables (Y_1, Y_2) instead of two events A and B. To be in line with the traditional notations, let $Y_1 = X$ and $Y_2 = \theta$. Also, for notational convenience, we will use $f(\cdot)$ for both discrete probability function (e.g., for discrete X, take $f(x) = P(X = x)$) and continuous pdf (e.g., for continuous X, take $f(x) = f_X(x)$).

Let $f(x)$ be a probability density of X, i.e., we have $f(x) \geq 0$ for all possible values of x and $\int f(x)dx = 1$ (if x only takes discrete values, replace the integral by the sum).

Suppose that $f(x, \theta)$ is the joint density, so that it takes values in both x and θ and has the properties that $f(x, \theta) \geq 0$ and $\iint f(x, \theta)\, dxd\theta = 1$. Note that

$$f(x) = \int f(x, \theta)\, d\theta \tag{8}$$

i.e., we obtain a marginal density $f(x)$ by integrating out the θ.

Now, suppose that we are interested in obtaining the conditional density $f(\theta|x)$. Then analogous to (3), we can write

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{f(x)} \tag{9}$$

(Note that (9) does not immediately follow from (3) since there are some results and gaps that must be filled in. Nevertheless, we see the exact match of (3) and (9), and so for this reason we also call (9) a form of Bayes theorem, applied to probability densities.)

Combining (9) with (8) and the fact that $f(x,\theta) = f(x|\theta)f(\theta)$, we obtain

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x,\theta)\,d\theta} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)\,d\theta} \tag{10}$$

but most times, we use the proportionality as in (7),

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta). \tag{11}$$

Mathematically, the equation (11) makes sense; as long as we have a conditional probability density $f(x|\theta)$ and a probability density $f(\theta)$, we can obtain the quantity $f(\theta|x)$. But how do we interpret the result for the statistical inference?

4 Bayesian Inference

Let us recall the equation (11). The primary variable of interest is θ. The θ is a random quantity that has a probability distribution. Without the data, we can only guess what the distribution of θ looks like, so we need to specify or "elicit" the **prior distribution** of θ, $f(\theta)$.

We then need to have data to come in to our model and inference. The quantity x can be thought of as data, which is actually of the form $x|\theta$, read x given θ. So θ does govern some mechanism in which the data x is generated. In a statistical problem, we are either given or need to specify the **likelihood** function $f(x|\theta)$.

Now, by (11), we then can obtain $f(\theta|x)$ by

$$f(\theta|x) \propto f(x|\theta)f(\theta).$$

Note that $f(\theta|x)$ is a function of θ but has the component x as well. The quantity $f(\theta|x)$ is interpreted as the updated distribution of θ *after* seeing the data x. Recalling that $f(\theta)$ is a distribution of θ prior to seeing the data (i.e., prior to the experiment), we now have the **posterior distribution** of θ as $f(\theta|x)$ which takes into an account the data x and which can be obtained only through $f(x|\theta)$ (likelihood) and $f(\theta)$ (prior) according to $f(\theta|x) \propto f(x|\theta)f(\theta)$.

The prior, likelihood, and posterior make up the core of the Bayesian analysis. The choice of prior and likelihood is up to the experimenter. If we have a "nice" form of both prior and likelihood, then the analysis becomes very simple and manageable.

5 Simple Cases

We illustrate the use of (11) by giving a couple of concrete examples involving known distributions.

5.1 Poisson-Gamma

Suppose that X, the data, is thought to have come from a Poisson distribution with an unknown parameter θ. Then we have the likelihood

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, 2, \ldots, \text{ and } \theta > 0. \tag{12}$$

In this case, the θ is the parameter of interest. If θ is thought to be distributed as a gamma distribution with known parameters α and β, then we have the prior (note the different representation of gamma pdf here)

$$f(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad \theta > 0,\ \alpha > 0,\ \beta > 0 \tag{13}$$

where $\Gamma(\alpha)$ is the gamma function satisfying $\Gamma(\alpha) = (\alpha - 1)!$.

Formally, we can write

$$\begin{aligned} X|\theta &\sim \text{Poisson}(\theta) \\ \theta &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

to signify that the likelihood $X|\theta$ has a Poisson distribution and the prior θ has a gamma distribution. Then it is natural to ask what the posterior $\theta|X$ looks like in this case.

For this, we use (11) with (12) and (13), obtaining

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)f(\theta) \\ &= \frac{e^{-\theta}\theta^x}{x!}\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha} \\ &\propto e^{-\theta}\theta^x\theta^{\alpha-1}e^{-\theta/\beta} \\ &= \theta^{\alpha+x-1}e^{-\theta((1+\beta)/\beta)} \end{aligned} \tag{14}$$

where the proportionality between the second and the third line of (14) is achieved by removing all the terms that does not involve θ. Then we have

$$f(\theta|x) \propto \theta^{\alpha+x-1}e^{-\theta((1+\beta)/\beta)} \tag{15}$$

where we can recognize the right-hand side of (15) as the numerator of the Gamma$(\alpha+x, \beta/(1+\beta))$ distribution.

So what we really have is

$$f(\theta|x) = \frac{\theta^{\alpha+x-1}e^{-\theta((1+\beta)/\beta)}}{\Gamma(\alpha+x)[\beta/(1+\beta)]^{(\alpha+x)}} = \frac{f(x|\theta)f(\theta)}{f(x)}$$

and it can be verified that the denominator

$$f(x) = \int_0^\infty f(x|\theta) f(\theta)\, d\theta = \frac{\Gamma(\alpha + x)[\beta/(1+\beta)]^{(\alpha+x)}}{x!\Gamma(\alpha)\beta^\alpha}$$

but the form of (14) or (15) is sufficient to recognize what the posterior should be.

Note that we can find $E(X)$ directly by

$$E(X) = \int x f(x)\, dx = \sum_{x=0}^{\infty} x \frac{\Gamma(\alpha + x)[\beta/(1+\beta)]^{(\alpha+x)}}{x!\Gamma(\alpha)\beta^\alpha} = \alpha\beta$$

but we can use

$$E(X) = E(E(X|\theta)) = E(\theta) = \alpha\beta$$

which is very convenient.

Similarly, we can obtain the variance of X by

$$\begin{aligned} \text{Var}(X) &= E(\text{Var}(X|\theta)) + \text{Var}(E(X|\theta)) = E(\theta) + \text{Var}(\theta) = \alpha\beta + \alpha\beta^2 \\ &= \alpha\beta(1+\beta) \end{aligned}$$

It is important to note that the posterior distribution $\theta|X$ *updates* the prior distribution θ with the data X. For out example, we started from the prior $\theta \sim \text{Gamma}(\alpha, \beta)$, but it was updated to the posterior $\theta|X \sim \text{Gamma}(\alpha + X, \beta/(1+\beta))$ with the addition of the data X. This will be a recurring theme.

In conclusion, from the prior and the likelihood

$$\begin{aligned} \theta &\sim \text{Gamma}(\alpha, \beta) \\ X|\theta &\sim \text{Poisson}(\theta) \end{aligned}$$

we obtain the posterior

$$\theta|X \sim \text{Gamma}(\alpha + X, \beta/(1+\beta))$$

for this problem.

5.2 Normal-Normal

We now suppose that we have a random sample $X_1, \ldots, X_n$ from normal (Gaussian) distribution with mean θ and variance σ^2. For convenience, let σ^2 be a known value and let θ be the parameter of interest. Then

$$p(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}, \quad -\infty < x_i < \infty,\ -\infty < \theta < \infty,\ \sigma > 0,\ i = 1, \ldots, n \tag{16}$$

and we can say that the data $X_i|\theta,\ i = 1, \ldots, n$ is from a normal distribution, and hence our likelihood is normal.

We then need to specify the prior distribution of θ. If we assume that θ also has a normal distribution, with known parameters mean μ and variance τ^2, then we get

$$f(\theta) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}, \quad -\infty < \theta < \infty, \ -\infty < \mu < \infty, \ \tau > 0 \tag{17}$$

and so we have a normal prior for θ

Or in other words, we have

$$\begin{aligned} X_i|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu, \tau^2) \end{aligned}$$

as likelihood and prior, respectively, and the quantities σ^2, μ, τ^2 are all known.

Notice that in (16), we have n of these functions that are independent, so that we work with the joint likelihood

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}. \tag{18}$$

Hence, if we use (11) again, with (18) and (17), we obtain

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|\theta) f(\theta) \\ &= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \right) \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \end{aligned} \tag{19}$$

To simplify this expression somewhat, we can use $\bar{x} = (x_1 + \cdots + x_n)/n$ rather than using all the data $x_1, \ldots, x_n$ (this is a use of sufficiency), and since we can show that $\bar{X}|\theta$ is $N(\theta, \sigma^2/n)$, we can rewrite (19) as

$$\begin{aligned} p(\theta|\bar{x}) &\propto p(\bar{x}|\theta) f(\theta) \\ &= \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} e^{-\frac{(\bar{x}-\theta)^2}{2\sigma^2/n}} \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \end{aligned}$$

and with some algebra, we can show that the posterior is of the form

$$\theta|\bar{X} \sim N\left(\frac{\sigma^2\mu + n\tau^2\bar{X}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} \right) \tag{20}$$

i.e., the posterior $\theta|\bar{X}$ is still normal with mean $\frac{\sigma^2\mu + n\tau^2\bar{X}}{\sigma^2 + n\tau^2}$ and variance $\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$. The expression for the mean and the variance of the posterior seems rather messy, but they do have a meaning. This is easier to see if we consider the precision (a reciprocal of variance) rather than the variance itself. For example, we can rewrite the posterior mean, $E(\theta|\bar{X}) = \frac{\sigma^2\mu + n\tau^2\bar{X}}{\sigma^2 + n\tau^2}$ as

$$\begin{aligned} E(\theta|\bar{X}) &= \frac{\sigma^2\mu + n\tau^2\bar{X}}{\sigma^2 + n\tau^2} \\ &= \frac{(\sigma^2\mu + n\tau^2\bar{X})/(\sigma^2\tau^2)}{(\sigma^2 + n\tau^2)/(\sigma^2\tau^2)} \\ &= \frac{\frac{1}{\tau^2}\mu + \frac{n}{\sigma^2}\bar{X}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\ &= \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{1}{(\sigma^2/n)}}\mu + \frac{\frac{1}{(\sigma^2/n)}}{\frac{1}{\tau^2} + \frac{1}{(\sigma^2/n)}}\bar{X} \end{aligned}$$

and so the posterior mean is the precision weighted average of the prior mean μ and the sample mean $\bar{X}$. And since the posterior variance is $\mathrm{Var}(\theta|\bar{X}) = \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}$, we have that the posterior precision is

$$\frac{1}{\mathrm{Var}(\theta|\bar{X})} = \frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2} = \frac{1}{\tau^2} + \frac{1}{(\sigma^2/n)}$$

i.e., the posterior precision is a sum of the prior precision and the data precision. Hence, both the posterior mean and the posterior variance (precision) comes out rather nicely.

More importantly, we see that the posterior is still normal so that we can easily do the analysis.

The examples of this section is what is known as **conjugate analysis**. This is because, if we have certain form of prior and likelihood, then we can automatically obtain a form of posterior.

However, there are only few valid conjugate prior-likelihood pairs, and in most real-life problems, we do not have this nice form. Fortunately, there are numerical methods which can approximate the posterior distribution. (So our situation is somewhat similar to differential equations where the exact solution is not always found but we can do some numerical approximation of a solution.)

The examples shown are called the **hierarchical models** because we have stages where we have $X|\theta$ then θ to build models and make inferences.