Bayesian Statistics

Anita Wang

2017 - 2018

Bayes' Rule

The **joint** probability mass or density function can be written as a product of the **prior distribution** $p(\theta)$ and **the sampling distribution** $p(y|\theta)$

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Bayes' Rule:

The **posterior** density

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

where $p(y) = \sum_{\theta} p(\theta)p(y|\theta)$ over all possible value of θ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ).

Bayes' Rule

An equivalent form omits the factor p(y), which does not depend on θ and, with fixed y, can thus be considered a constant, yielding the **unnormalized posterior density**,

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The second term in this expression, p(y|θ), is taken here as a function of θ, not of y.

Prediction

- **Prior predictive distribution** (also called marginal distribution of y)

$$p(y) = \int p(y,\theta)d\theta = \int p(\theta)p(y|\theta)d\theta$$

prior because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

Prediction

- Posterior predictive distribution

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\
&= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\
&= \int p(\tilde{y}|\theta) p(\theta|y) d\theta.
\end{aligned}
$$

Once the data y have been observed, the unknown observable $\tilde{y}$ can be predicted. For example, $y = (y_1, y_2, \ldots, y_n)$ may be the vector of recorded weights of an object weighed n times on a scale, $\theta = (\mu, \sigma^2)$ is the prior, and $\tilde{y}$ may be the yet to be recorded weight of the object in a planned new weighing.

Likelihood

Using Bayes' rule with a chosen probability model means that the data y affect the posterior inference only through $p(y|\theta)$. $p(y|\theta)$ is regarded as a function of θ for fixed y is **likelihood function**.

Likelihood principle

for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ.

Likelihood and odds ratios

- The ratio of the posterior density $p(\theta|y)$ evaluated at the points θ_1 and θ_2 under a given model is called the posterior odds for θ_1 compared to θ_2.

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)}$$

The posterior odds are equal to the prior odds multiplied by the likelihood ratio $p(y|\theta_1)/p(y|\theta_2)$.

Example

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is rare, since the frequency of occurrence of the gene is low in human populations.

Prior distribution

- Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one 'good' and one 'bad' hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene.
- Set: to be a carrier of the gene ($\theta = 1$) or not ($\theta = 0$)

- The prior distribution for unknown θ is

$$\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}$$

Likelihood

- The data used to update the prior information consist of the affection status of the woman's sons. Suppose she has two sons, neither of whom is affected. The outcomes of the two sons are exchangeable independent
- Let $y_i = 1$ or 0 denote an affected or unaffected son.

- The likelihood function is:

$$\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 1) = (0.5)(0.5) = 0.25$$

$$\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 0) = (1)(1) = 1$$

Posterior distribution

- In particular, interest is likely to focus on the posterior probability that the woman is a carrier.

$$\begin{aligned}\Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 1)\Pr(\theta = 1) + p(y|\theta = 0)\Pr(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20.\end{aligned}$$

- Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' rule provides a formal mechanism for determining the extent of the correction.

Prior and Posterior Odds

- The prior odds of the woman being a carrier

$$\frac{p(\theta = 1)}{p(\theta = 0)} = 0.5/0.5 = 1$$

- The likelihood ratio is

$$\frac{\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 1)}{\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 0)} = 0.25/1 = 0.25$$

- The posterior odds is $1(0.25) = 0.25$
- Conversely, the posterior probability

$$p(\theta = 1|y) = 0.25/(1+0.25) = 0.2$$

Adding more data

- A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed.
- For example, suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution

$$\Pr(\theta = 1 | y_1, y_2, y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2) + (1)(0.8)} = 0.111$$

Example

- Spelling correction is an important technique. Suppose someone types ‘radom.’ How should that be read? It could be a misspelling or mistyping of ‘random’ or ‘radon’ or some other alternative, or it could be the intentional typing of ‘radom’ (as in its first use in this paragraph).

$$\Pr(\theta|y = \text{‘radom’}) \propto p(\theta)\Pr(y = \text{‘radom’}|\theta)$$

$$p(\text{random}|\ \text{‘radom’}) = \frac{p(\theta_1)p(\text{‘radom’}|\theta_1)}{\sum_{j=1}^{3} p(\theta_j)p(\text{‘radom’}|\theta_j)}$$

$$\theta_1 = \text{random},\ \theta_2 = \text{radon},\ \theta_3 = \text{radom}$$

Prior distribution

- The prior probabilities $p(\theta_j)$ can most simply come from frequencies of these words in some large database

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

- For the documents that we encounter, the relative probability of 'radom' seems much too high. We looked up the word in Wikipedia and found that it is a medium-sized city, home to 'the largest and best-attended air show in Poland . . .'
- We may have prior information or beliefs that have not yet been included in the model.

Likelihood

- Here are some conditional probabilities from Google's model of spelling and typing errors:

θ	$p(\text{'radom'}\|\theta)$
random	0.00193
radon	0.000143
radom	0.975

- 97% chance that this particular five-letter word will be typed correctly, a 0.2% chance of obtaining this character string by mistakenly dropping a letter from 'random,' and a much lower chance of obtaining it by mistyping the final letter of 'radon.'

Posterior distribution

- We multiply the prior probability and the likelihood to get joint probabilities and then renormalize to get posterior probabilities:

θ	$p(\theta)p(\text{`radom'}\|\theta)$	$p(\theta\|\text{`radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

- the typed word ‘radom’ is about twice as likely to be correct as to be a typographical error for ‘random,’ and it is very unlikely to be a mistaken instance of ‘radon.’

Decision making, model checking and improvement

- The first approach is to accept that the word was typed correctly
- The second option would be to question this probability by saying, for example, that ‘radom’ looks like a typo and that the estimated probability of it being correct seems much too high.
- Questioning the prior: The prior probabilities, on the other hand, are highly context dependent. The word ‘random’ is of course highly frequent in our own writing on statistics, ‘radon’ occurs occasionally, while ‘radom’ was entirely new to us.
- Label x as the contextual information

$$p(\theta|x, y) \propto p(\theta|x)p(y|\theta, x).$$

Thanks