Part TWO

Mathematical Methods for Physics

from **On the Studies of Physics and Her Axillary Studies** by Shing Hin (John) Yeung

Chapter 36

Bayesian Inference: Single Parameter Models

36.1 Estimate Priors using Unnormalised Density

In a single parameter model, we only worried about one of the parameters, such as the probability of "Heads" in each tossing. There should be other parameters as factors, however, we have supposed them as fixed. Under this assumption, we may manipulate the definition of posterior probability eq. (35.7), while the parameters of the evidence (i.e. in the denominator) becomes fixed in deducing the likelihood[1]. Hence,

36.1.1 Theorem (Unnormalised Density)

A statistical model which consists of a probability density function $P(x \mid \theta)$ in which x is the random variable and θ are any arbitrary and particular parameters. The unnormalised density remains the product of likelihood function $P(\theta \mid x)$ and the prior $P(x)$. So that other parameters, which supposed to be fixed, become a constant. Which is

$$P(x \mid \theta) \propto P(\theta \mid x) \times P(x) \tag{36.1}$$

[1] Hence when using the method of maximum likelihood, any fixed constants are cancelled.

We may use this for estimating a prior for binomial distributions. This is useful in the world where two (i.e. 2) options are available for each run.

Exercise 36.1.1: Estimate a Prior for Coin Tossing Experiments

A coin tossing experiment which has x successes with individual probability $p = 0.5$ for, say ten (i.e. 10) tosses. Find the prior distribution $f(x)$.

Solution:

In this problem, we knew that the number of successes x is the variable, and it determines the prior which is the number of successes without any effective parameters. We also knew that the posterior distribution, which is

$$f(x \,|\, p = 0.5, n = 10) = \binom{10}{x} 0.5^x 0.5^{10-x} \tag{36.2}$$

The problem asked for a distribution that is not affected by any other parameters. First of all, we are extending the single-valued probability into its distribution form. Hence a prior of x, by means of a probability which has not encountered for other conditions, such as total number of tosses and probability of each toss. Using the unnormalised density eq. (36.1), which is

$$f(x \,|\, 0.5,\, 10) \propto f(n, p \,|\, x) \times f(x) \tag{36.3}$$

Thus we ought to find the functions determined by all of the parameters and the random variable x, and delete any constants from eq. (36.2). From eq. (36.3), the function that is determined by all variables is the latter part. Which is

$$f(10, 0.5 \,|\, x) \propto 0.5^x 0.5^{10-x} = 0.5^{10} \tag{36.4}$$

hence the prior should have the same form to obtain the same function of posterior distribution. Equation (36.4) may apply to the prior. The remaining term in eq. (36.2) which is

$$f(x) \propto \binom{10}{x} \tag{36.5}$$

is not applicable since it is irrelevant to the parameters for conditional probability.

36.2 Informative Prior

We should not assume that we always know the exact prior from our analysis, hence finding prior becomes very important in Bayesian inferences. In the case that the posterior distribution is found, such as in coin tossing experiments one may have the insight of "Heads" and "Tail" are distributed by binomial distributions. In Exercise 36.1.1 we have distilled the prior by means of unnormalised

density. For this time, we make generalisation of the coin-tossing case and we resumed the notations of p as the probability and n as the total times of coin tossing. So that the prior, as per Exercise 36.1.1 states is

$$P(n, p \,|\, x) \propto p^x \, (1 - p)^{n - x} \tag{36.6}$$

The next stage is which distribution will fit this. In here it is the matter of inductions and testing your inductions of the model. One statistical model of the prior could be the beta distribution $B(\alpha, \beta)$. There are two (i.e. 2) parameters for the prior distribution α and β, which is namely the **hyperparameters**. It is then said the prior distribution, in this case, is indexed by both α and β. So the prior distribution has the form

$$P(n, p \,|\, x) \propto p^{(\alpha - 1)} \, (1 - p)^{(\beta - 1)} \tag{36.7}$$

We may substitute this into the posterior distribution and make such controlled by the hyperparameters. Hence

$$\begin{aligned} P(x \,|\, n, p) &\propto p^x \, (1 - p)^{(n - x)} \, p^{(\alpha - 1)} \, (1 - p)^{(\beta - 1)} \\ &= p^{(x + \alpha - 1)} \, (1 - p)^{(n + \beta - 1)} \end{aligned} \tag{36.8}$$

We may say the posterior distribution follows a beta distribution $B(p \,|\, x + \alpha, n + \beta)$, provided that we knew the exact value of both α and β. The saying of the posterior originally follows the binomial distribution, now also the beta distribution. Meaning that the beta distribution is **conjuagcy** of the binomial distribution. Such definition is:

36.2.1 Definition (Conjugacy)

Suppose a class of sampling distributions $\mathcal{S}(x \,|\, \theta)$, and $\mathcal{P}$ is a class of prior distributions. Class $\mathcal{P}$ is conjugate to $\mathcal{S}$ if $P(\theta \,|\, x) \in \mathcal{P}$, which for all of the $P(\cdot \,|\, \theta) \in \mathcal{S}$ and $\mathcal{S}$ if $P(\cdot) \in \mathcal{P}$.

In here we follow Gelman et al. (2004) in which the conjugacy is a natural conjugate and the prior has the same form as the likelihood function.

Conjugate prior distributions, such as above, makes computation easier. We may put them in analytical forms, and they are often in good approximations. Conjugate makes the posterior distribution understandable which makes readers understand how the hyperparameters would operate upon the model. In more complicated models, conjugate distributions may not be found but non-conjugate distributions.

36.3 Non-informative Prior is a Backup when Prior is Unknown

This chapter focuses on the prior and how to make the posterior relevant. Prior is important in Bayesian inference which affects the posterior probability. Laplace (Syversveen 1998, p.1 with amendments due to notations in below) has stated the following:

> When nothing is known about $\boldsymbol{\theta}$ in advance, let the prior $P(x)$ be a uniform distribution, that is, let all possible outcomes of $\boldsymbol{\theta}$ have the same probability.

This is called the **Principle of Insufficient Reason**. The most obvious application would be you don't know what is the distribution of obtaining heads in coin tossing experiments. Since there are only two (i.e. 2) outcomes, then assume the experiment is fair hence the probability to obtain for each outcome is half (i.e. 0.5).

In fact, one should avoid using Principle of Insufficient Reason, which is assigning uniform probabilities into the prior. For this, a philosophical approach should be done. Suppose an apparatus produces systematic error which is waiting to be investigated. Should I be assuming the probability of the next person which produces the systematic error due to that equipment is the same? In terms of the probability which isolates any conditions (or causes), it is plausible. However, assigning the same probability for all random variables (in here the users in sequence) will not help to find out what is the problem. Hence, there is a long debate for not to use uniform distribution as prior, and how to encounter a better form of prior.

36.3.1 Jeffrey Prior

This method was described in Jeffreys (1946, cited in Syversveen 1998) in which Jeffrey Prior is proportional to the square root of Fisher Information Matrix. Which is

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)} \tag{36.9}$$

where Fisher Information Matrix $\mathcal{I}(\theta)$ is defined as

$$\mathcal{I}(\theta) \equiv \left\langle \left(\frac{\partial \ell}{\partial \theta} \right)^2 \mid \theta \right\rangle \tag{36.10}$$

and ℓ is the log-likelihood function.

36.4 Summarising Posterior Distribution

For practical purposes and for future computations, posterior distributions can be summarised in terms of:

- mean value
- median
- mode
- standard deviation
- interquartile range

Where the first three (i.e. 3) items are summarising the location parameters, and the latter two (i.e. 2) refers to the scale parameter. In the case when the posterior distribution is in closed form, the values should be summarised easily.

The mean value can be interpreted as the expected value. In the long run of experiments or using frequency approach of probability, this is the mode as well. The mode can be interpreted as the most probable random variable, provided by the data (or model). While finding mode value is much easier[2] than finding mean or median values, hence it is heavily used in summaries.

Bibliography

Gelman, A. et al. (2004). **Bayesian Data Analysis**. en. 2nd ed. CRC Press.

Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems". en. In: **Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences** 186.1007, pp. 453–461. DOI: 10.1098/rspa.1946.0056.

Syversveen, A. (1998). **Noninformative Bayesian Priors. Interpretation And Problems With Construction And Applications.** en. [Online; accessed 20-January-2017]. URL: http://www.ime.unicamp.br/~veronica/MI402/Randi21998.pdf.

Appendix 36.A Working Diary

36.A.1 19/01/2017

Started this chapter today.

Finish this chapter tmr by non-informative priors and vague priors as well. Will continue this chapter then if there is more contents.

[2] It is the largest probability value one obtained.

36.A.2 23/01/2017

Working on sections 36.2 and 36.4.

Appendix 36.B Selected Readings and their Reviews

Appendix 36.C Connections to Other Topics

Appendix 36.D Documentations