MAST90125: Bayesian Statistical learning

Lecture 2: Prior choice

Prepared by Feng Liu and Guoqi Qian

Priors

The posterior distribution $p(\theta|y)$ can be written as follows,

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

- One of the underlying differences between Bayesian and frequentist statistics is that in Bayesian statistics, you deal with a full probability model, that is a probability model for both data (y) and parameters (θ).

Priors

The posterior distribution $p(\theta|y)$ can be written as follows,

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

- One of the underlying differences between Bayesian and frequentist statistics is that in Bayesian statistics, you deal with a full probability model, that is a probability model for both data (y) and parameters (θ).
  - In practice, this means we need to specify a prior distribution for the set of parameters before starting analysis.
  - Moreover, it means $p(y|\theta)$, which in frequentist statistics is viewed as the likelihood $L(\theta)$, is now the sampling distribution conditional on parameters θ.
- This lecture will focus on the impact the choice of prior has on inference.

Practical note

- In many applications, it is more convenient to ignore the normalising constant $p(y)$ and write
$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \propto p(y|\theta)p(\theta).$$
- More generally, this dropping out often extends to the normalising constant of the likelihood and prior, so that Bayesian statisticians pay particular attention to distribution kernels. Some examples of kernels include:
  - Gamma distribution: $x^{\alpha-1}e^{-\beta x}$
  - Beta distribution: $p^{\alpha-1}(1-p)^{\beta-1}$
  - Normal distribution if μ is only unknown: $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- That is for our purposes, if we have a joint distribution $f(y, \theta)$, which can be written as $h(y)g(y, \theta)$, then the distribution kernel is $g(y, \theta)$.

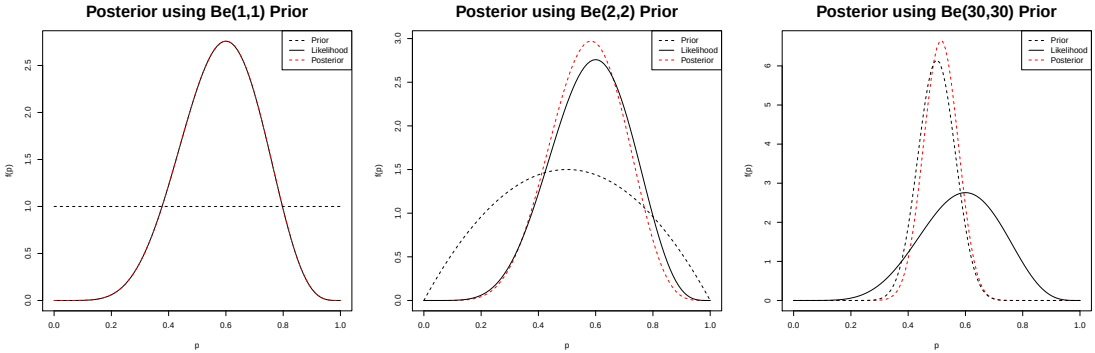
Impact of prior choice on posterior: Binomial example

- The requirement for a full probability model means the posterior probability can be viewed as a weighted average on prior with weights proportional to the likelihood?

Impact of prior choice on posterior: Binomial example

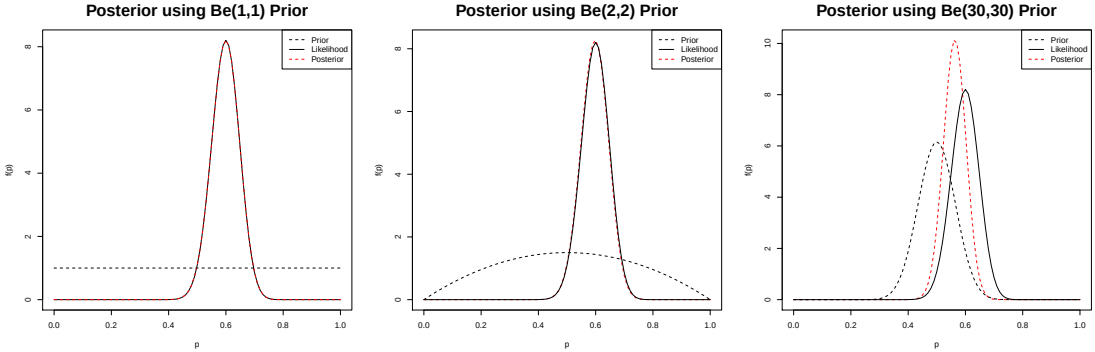
- The requirement for a full probability model means the posterior probability can be viewed as a weighted average on prior with weights proportional to the likelihood? $\int_\theta p(y|\theta)p(\theta)d\theta$.
- To demonstrate this, consider the estimation of the probability parameter p (i.e. θ), when we have observed 6 successes and 4 failures.
- We suspect *a priori* that p is drawn from a beta ($Be(\alpha, \beta)$) distribution, such that $E(p) = 0.5$.

Impact of prior choice on posterior



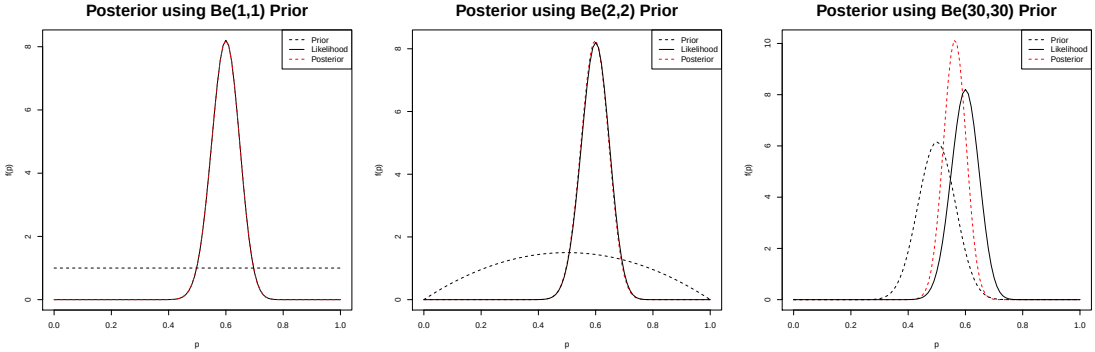
- Notice how the choice of α, β controls the relative weight of likelihood and prior in the posterior distribution.
- But what would happen if we observed 60 successes and 40 failures?

Impact of prior choice on posterior



- Increasing the sample size increased the weight placed on the likelihood when constructing the posterior.

Impact of prior choice on posterior



- Increasing the sample size increased the weight placed on the likelihood when constructing the posterior. Why does (60,40) case have such an impact?

The informative prior

- Many real-world examples of priors used in the binomial example would be described as 'informative'.
  - That is, the prior conveyed information in addition to the likelihood that influenced the posterior distribution.
- However two questions arise?
  - In the example above which prior would be 'least' informative?
  - More importantly, why would you want to use an informative prior?

Why use an informative prior

- From a frequentist perspective, the use of an informative prior may seem like the researcher is planning to bias results.
- The appeal of an informative prior may become more apparent if we consider the following scenario.
  - Let's say two random samples $\mathbf{y}_1, \mathbf{y}_2$, where observations are, conditional on θ, *i.i.d* are drawn from the same model. We want to make inference about the model parameters $\boldsymbol{\theta}$.
- As the samples are *i.i.d* and drawn from the same model, we can combine the two samples (that is $\mathbf{y} = (\mathbf{y}_1 \ \ \mathbf{y}_2)$), and calculate the posterior as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}.$$

Why use an informative prior

- But the posterior can also be written as,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}_2|\boldsymbol{\theta})p(\mathbf{y}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}_2, \mathbf{y}_1)} = \frac{p(\mathbf{y}_2|\boldsymbol{\theta})p(\mathbf{y}_1, \boldsymbol{\theta})}{p(\mathbf{y}_2, \mathbf{y}_1)} = \frac{p(\mathbf{y}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_1)p(\mathbf{y}_1)}{p(\mathbf{y}_2, \mathbf{y}_1)}$$
$$= \frac{p(\mathbf{y}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_1)}{p(\mathbf{y}_2|\mathbf{y}_1)},$$

- such that we can calculate the posterior from the first sample only, $p(\boldsymbol{\theta}|\mathbf{y}_1)$, and then use $p(\boldsymbol{\theta}|\mathbf{y}_1)$ as a prior when analysing the second sample.

- Hence you can think of the informative prior as representing the information gained from all past experiments of the same phenomena.

The uninformative prior

- In a situation where little or nothing is known about the underlying mechanism, a researcher would likely prefer a prior that would convey minimal information.
- Thinking back to the binomial example, which prior would appear to convey least information.
  - The $Be(1,1)$ prior $\Leftrightarrow$ uniform. This is referred to as a flat prior and implies that the posterior and likelihood share the same kernel.

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

  - But does this imply that Bayesian inference with a flat prior is equivalent to inference based on Maximum likelihood?

Proper vs improper

- When we have discussed priors, we have referred to the prior distribution.
- However so far, we have not checked if the prior we specify is a valid distribution, that is, among other conditions, satisfies

$$\int p(\theta) d\theta = 1.$$

- But to obtain a valid or proper posterior, is it necessary for the prior to be a valid distribution, that is proper?
- To find an answer, consider the binomial likelihood with a beta $\text{Be}(\alpha, \beta)$ prior.

Proper vs improper

- If we look at the joint distribution $p(y, \theta) = \Pr(y|\theta)p(\theta) =$

Proper vs improper

- If we look at the joint distribution $p(y, \theta) = \Pr(y|\theta)p(\theta) =$

$$\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

  we see that $\theta|y \sim \text{Be}(\alpha + y, \beta + n - y)$.
- But what is the support for the parameters of a Beta distribution?

Proper vs improper

- If we look at the joint distribution $p(y, \theta) = \Pr(y|\theta)p(\theta) =$

$$\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

  we see that $\theta|y \sim \text{Be}(\alpha + y, \beta + n - y)$.
- But what is the support for the parameters of a Beta distribution?
  - The strictly positive real numbers.
- So $\alpha + y > 0$ and $\beta + n - y > 0$ must hold for a proper posterior, but there is no requirement for $\alpha > 0$ or $\beta > 0$ so the prior need not be proper.

Proper vs improper

- If we look at the joint distribution $p(y, \theta) = \Pr(y|\theta)p(\theta) =$

$$\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

  we see that $\theta|y \sim \text{Be}(\alpha + y, \beta + n - y)$.
- But what is the support for the parameters of a Beta distribution?
  - The strictly positive real numbers.
- So $\alpha + y > 0$ and $\beta + n - y > 0$ must hold for a proper posterior, but there is no requirement for $\alpha > 0$ or $\beta > 0$ so the prior need not be proper. The posterior distribution is valid in nature. We do not need prior to be proper. $\int p(\theta|y)d\theta = 1$.

Conjugacy

- In the previous example, we showed that if we assume a Beta prior in conjunction with a binomial likelihood for the proportion θ, then the posterior would also be beta distributed.
- This situation, where the prior and posterior are of the same distribution, but with different parameters is when the prior can be referred to as a 'conjugate prior'.
- Note the statement on conjugacy also depends on the likelihood and requires the likelihood to be of the same functional form as the prior.

Conjugacy and exponential families

- From previous subjects, you may be familiar with exponential families. The pdf of distributions belonging to the exponential family can be written

$$p(y|\boldsymbol{\theta}) = f(y)g(\boldsymbol{\theta})e^{\eta(\boldsymbol{\theta})'u(y)},$$

where $\eta(\boldsymbol{\theta})$ is a vector of natural parameters. Furthermore $\eta(\boldsymbol{\theta})$ and $u(y)$ have the same length.

- You may also be familiar that if $y_i \sim F$ where F is a distribution in the exponential family and y_i are $i.i.d.$, then the likelihood can be written as,

$$\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) = \left\{ \prod_{i=1}^{n} f(y_i) \right\} g(\boldsymbol{\theta})^n e^{\eta(\boldsymbol{\theta})'\{\sum_{i=1}^{n} u(y_i)\}},$$

where the vector $\sum_{i=1}^{n} u(y_i)$ denotes the set of minimal sufficient statistics.

Conjugacy and exponential families

- If we assume the prior has the form

$$p(\boldsymbol{\theta}) = g(\boldsymbol{\theta})^{\kappa} e^{\eta(\boldsymbol{\theta})'\boldsymbol{\nu}},$$

then it can be deduced that the posterior

$$\begin{aligned} p(\boldsymbol{\theta}|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= g(\boldsymbol{\theta})^n e^{\eta(\boldsymbol{\theta})'\{\sum_{i=1}^n u(y_i)\}} \times g(\boldsymbol{\theta})^{\kappa} e^{\eta(\boldsymbol{\theta})'\boldsymbol{\nu}} \\ &= g(\boldsymbol{\theta})^{n+\kappa} e^{\eta(\boldsymbol{\theta})'\{\sum_{i=1}^n u(y_i)+\boldsymbol{\nu}\}}, \end{aligned}$$

has the same functional form as the prior and likelihood and thus is conjugate.

Conjugacy and exponential families

- If we assume the prior has the form

$$p(\boldsymbol{\theta}) = g(\boldsymbol{\theta})^{\kappa} e^{\eta(\boldsymbol{\theta})'\boldsymbol{\nu}},$$

then it can be deduced that the posterior

$$\begin{aligned} p(\boldsymbol{\theta}|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= g(\boldsymbol{\theta})^n e^{\eta(\boldsymbol{\theta})'\{\sum_{i=1}^n u(y_i)\}} \times g(\boldsymbol{\theta})^{\kappa} e^{\eta(\boldsymbol{\theta})'\boldsymbol{\nu}} \\ &= g(\boldsymbol{\theta})^{n+\kappa} e^{\eta(\boldsymbol{\theta})'\{\sum_{i=1}^n u(y_i)+\boldsymbol{\nu}\}}, \end{aligned}$$

has the same functional form as the prior and likelihood and thus is conjugate.

- It turns out that distributions in the exponential family are the only distributions that have natural conjugate prior distributions.

Jeffreys' prior

- To concludes this lecture, consider the question, are flat priors the only choice for an 'uninformative' prior?
- Jeffreys proposed a non-informative prior based on considering one-to-one transformations of θ, $h(\theta) = \phi$. If $p(\theta)$ is the prior for θ, then

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1}.$$

- Jeffreys' invariance principle leads to a non-informative prior defined as,

$$p(\theta) \propto \sqrt{J(\theta)}, \quad \Leftrightarrow \quad p(\phi) \propto \sqrt{\frac{d\theta}{d\phi} J(\theta) \frac{d\theta}{d\phi}} = \sqrt{J(\phi)}.$$

  Fisher information: $J(\theta) = -E\left(\frac{d^2 \log(p(y|\theta))}{d\theta^2}|\theta\right)$ & $J(\phi) = -E\left(\frac{d^2 \log(p(y|\theta))}{d\phi^2}|\theta\right)$.

- To conclude this lecture, lets consider an example.

Jeffreys prior: Example

- Consider the case of a Poisson likelihood

$$\Pr(y_1, \dots y_n|\lambda) = \prod_{i=1}^{n} \Pr(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}$$

- Quiz: Now determine Jeffreys' prior.

Jeffreys prior: Example

- Consider the case of a Poisson likelihood

$$\Pr(y_1, \dots y_n|\lambda) = \prod_{i=1}^{n} \Pr(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}$$

- Quiz: Now determine Jeffreys' prior.

$$\log(\Pr(\mathbf{y}|\lambda)) = c + n\bar{y}\log(\lambda) - n\lambda$$

$$\frac{d\log(\Pr(\mathbf{y}|\lambda))}{d\lambda} = \frac{n\bar{y}}{\lambda} - n$$

$$\frac{d^2\log(\Pr(\mathbf{y}|\lambda))}{d\lambda^2} = -\frac{n\bar{y}}{\lambda^2}$$

$$E\left(\frac{d^2\log(\Pr(\mathbf{y}|\lambda))}{d\lambda^2}\right) = -\frac{n\lambda}{\lambda^2} = -\frac{n}{\lambda}$$

Hence $p(\lambda) \propto 1/\sqrt{\lambda} = \lambda^{0.5-1}e^{-0\times\lambda} = \text{Ga}(0.5, 0)$, an improper prior.