STATS 700 Bayesian Inference and Computation **Fall 2017**

Lecture 3: Sept 18, 20 Single-parameter model

Lecturer: CHEN Yang *Scribe: Yuequan Guo, Yuanzhi Li, Zoey Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Comparison between MLE and Bayesian inference

- Frequentist and Bayesian inferences might converge as sample sizes increase.
- The choice of prior in Bayesian inference has significant implication for the inference especially for a small sample size

3.2 Binomial model

data: Bernoulli/ Binomial
prior: Beta / Uniform
posterior: (Unnormalized) Beta

Binomial model: $X_1, X_2, \ldots, X_n \mid p \sim B(n, p)$. We assume the prior for p follows Beta(α, β).

$$\begin{aligned} \text{Likelihood}: \quad & \mathrm{P}(X_1, \ldots, X_n \mid p) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}, \\ \text{Prior}: \quad & \mathrm{P}(p) \propto p^{\alpha-1}(1-p)^{\beta-1}. \end{aligned}$$

Since the parameters (α, β) in the prior is considered fixed, we can omit the normalizing constant in the prior. The posterior is the product of the likelihood and the prior.

$$\begin{aligned} \text{Posterior}: \quad & \mathrm{P}(p \mid X_1, \ldots, X_n) = \mathrm{P}(p) \cdot \mathrm{P}(X_1, \ldots, X_n \mid p) \\ & \propto p^{\alpha-1+\sum_{i=1}^n X_i}(1-p)^{\beta-1+n-\sum_{i=1}^n X_i} \end{aligned}$$

Hence, the posterior would be Beta$(\alpha + \sum_{i=1}^n X_i,\ \beta + n - \sum_{i=1}^n X_i)$. Note that if $\alpha = \beta = 1$, the prior is simply uniform distribution on $(0, 1)$, while the posterios will be Beta$(\sum_{i=1}^n X_i + 1,\ n - \sum_{i=1}^n X_i + 1)$.

3.3 Gaussian model

3.3.1 Gaussian with unknown mean but known variance

data: normal
prior: normal
posterior: normal

Gaussian model with unknown mean and known variance: $X_1, X_2, \ldots, X_n \mid \theta \sim \mathcal{N}(\theta, \sigma_0^2)$. We assume the prior for the mean θ also follow a Gaussian distribution $\mathcal{N}(\mu_0, \tau_0^2)$.

$$\begin{aligned}
\text{Likelihood}: &\quad \mathrm{P}(X_1, \ldots, X_n \mid \theta) \propto \exp\left\{\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i - \theta)^2\right\} \\
\text{Prior}: &\quad \mathrm{P}(\theta) \propto \exp\left\{\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} \\
\text{Posterior}: &\quad \mathrm{P}(\theta \mid X_1, \ldots, X_n) \propto \exp\left\{\left(\frac{n}{2\sigma_0^2} + \frac{1}{2\tau_0^2}\right)\theta^2 - \left(\frac{n\overline{X}}{\sigma_0^2} + \frac{1}{\tau_0^2}\right)\theta + C\right\}, \text{ for some constant C.}
\end{aligned}$$

Hence, the posterior for θ is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with:

$$\sigma^2 = \frac{1}{n/\sigma_0^2 + 1/\tau_0^2}, \quad \mu = \frac{\frac{n\overline{X}}{\sigma_0^2} + \frac{\mu_0}{\tau_0^2}}{n/\sigma_0^2 + 1/\tau_0^2}.$$

We can see that the variance for the posterior satisfy:

$$\frac{1}{\sigma^2} = \frac{n}{\sigma_0^2} + \frac{1}{\tau_0^2}. \tag{3.1}$$

Note that for any normal distribution $\mathcal{N}(\mu, \sigma^2)$, the Fisher information for the mean μ is $\mathcal{I}(\mu) = 1/\sigma^2$. Equation (**??**) indicates that the posterior fisher information is the sum of the information from prior and the information from the likelihood. Also, the posterior mean is a weighted average of the mean from prior and likelihood, and the weights are proportional to Fisher information.

- The inverse of the variance is the precision (i.e. the information).
- The posterior info is the addition of the info of the prior and the data.
- The posterior mean is weighted average of the prior and the data and the weights are proportional to information.

shrinkage estimators The posterior is shrinking towards the prior and the data depending on the relative information

posterior predictive distribution

With the posterior, assume we want to predict the distribution for some new data $\tilde{y}$. From

$$\mathrm{P}(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y) \mathrm{d}\theta,$$

we can see that the predictive posterior for $\tilde{y}$ is also normal. Below we calculate the mean and variance for this predictive posterior.

Recall: $\mathrm{E}(\tilde{y} \mid \theta) = \theta$, $\mathrm{Var}(\tilde{y} \mid \theta) = \sigma^2$. Then we have:

$$\begin{gathered}
\mathrm{E}(\tilde{y} \mid y) = \mathrm{E}\left[\mathrm{E}(\tilde{y} \mid \theta, y) \mid y\right] = \mu, \\
\mathrm{Var}(\tilde{y} \mid y) = \mathrm{E}[\mathrm{Var}(\tilde{y} \mid \theta, y) \mid y] + \mathrm{Var}[\mathrm{E}(\tilde{y} \mid \theta, y) \mid y] = \mathrm{E}(\sigma_0^2 \mid y) + \mathrm{Var}(\theta \mid y) = \sigma_0^2 + \sigma^2.
\end{gathered}$$

3.3.2 Gaussian with known mean but unknown variance

data: Normal

$$p(y_i : 1 \leq i \leq n | \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2) = (\sigma^2)^{-\frac{n}{2}} \exp(-\frac{nv}{2\sigma^2})$$

where $v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$.

prior: Inverse Gamma

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\frac{\beta}{\sigma^2})$$

where (α, β) are hyperparameters of the inverse-gamma distribution.

posterior: Inverse Gamma

$$p(\sigma^2 | y_i : 1 \leq i \leq n) \propto (\sigma^2)^{-(\alpha+\frac{n}{2}+1)} \exp(-\frac{1}{\sigma^2}(\beta + \frac{nv}{2}))$$

which is also an Inverse-Gamma distribution with parameters $(\alpha + \frac{n}{2}, \beta + \frac{nv}{2})$.

3.4 Poisson model

data: Poisson (number of counts)

$$p(y_i : 1 \leq i \leq n | \theta) \propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}$$

prior: Gamma

$$p(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}$$

where (α, β) are hyperparameters of the prior Gamma distribution.

posterior: Gamma

$$p(\theta | y_i : 1 \leq i \leq n) \propto e^{-(\beta+n)\theta}\theta^{(\sum_{i=1}^{n} y_i + \alpha - 1)}$$

which is also a Gamma distribution with parameters $(\sum_{i=1}^{n} y_i + \alpha, \beta + n)$.

prior predictive distribution: Negative Binomial
The prior predictive distribution of a single observation from the Poisson model is

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)} = \frac{\Gamma(\alpha + y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}} = \binom{\alpha + y - 1}{y}(\frac{\beta}{\beta+1})^{\alpha}(\frac{1}{\beta+1})^{y}$$

which is a negative binomial distribution with parameters (α, β).

3.5 Exponential model

data: Exponential (length of waiting time)

$$p(y_i : 1 \leq i \leq n | \theta) \propto \theta^n \exp(-\theta \sum_{i=1}^{n} y_i)$$

prior: Gamma

$$p(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}$$

which is also a Gamma distribution with parameters $(\sum_{i=1}^{n} y_i + \alpha, \beta + n)$.

posterior: Gamma

$$p(\theta|y_i : 1 \leq i \leq n) \propto e^{-(\beta+\sum_{i=1}^{n} y_i)\theta}\theta^{(n+\alpha-1)}$$

3.6 Discussion on posterior distribution

- The posterior distribution is centered around a point that represents a compromise between the prior and the data and the compromise is controlled to a greater extent by the data as the sample size increases
- The posterior variance is on average smaller than the prior variance
- The posterior will become more concentrated if the prior and the data agrees
- The posterior will become diffuse if they conflict. If this happens, this implies inconsistency between the prior and the data.

3.7 Informative prior distribution

- Population interpretation
- State of knowledge interpretation
- Cover all possible values

3.8 Summary statistics of posterior distribution

- mean, median, mode, interquartile, etc.
- standard deviation
- $100(1-\alpha)\%$ central posterior interval vs $100(1-\alpha)\%$ highest posterior density region. (See page 33 Figure 2.2 from [**?**])

References

[1] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press, 2012.

[2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, Volume 2. CRC press Boca Raton, FL, 2014.