Bayesian Essentials

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Most of the contents are from the Bayesian course taught by Mattias Villani

http://www.mattiasvillani.com/teaching/bayesian-statistics/

The Likelihood Function

- EXAMPLE (BERNOULLI).

$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta).$$

- Likelihood:

$$\begin{aligned} p(x_1, ..., x_n|\theta) &= p(x_1|\theta) \cdots p(x_n|\theta) \\ &= \theta^s(1-\theta)^f, \end{aligned}$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes in the Bernoulli trials and $f = n - s$ is the number of failures.

- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.

Learning From Data - Bayes' Theorem

- Given that you have formulated a distribution for θ, $p(\theta)$, how can we learn from data? That is, how do we make the transition from $p(\theta) \rightarrow p(\theta|Data)$? Bayes' theorem is the key.

- One form of Bayes' theorem reads (A and B are events)

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

  So that Bayes' theorem 'reverses the conditioning', i.e. takes us from $p(B|A)$ to $p(A|B)$.

- Let $A = \theta$ and $B = Data$

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

- Interpreting the likelihood function as a probability density for θ is just as wrong as ignoring the factor $p(A)/p(B)$ in Bayes' theorem.

Bayesian updating

- Suppose: you already have $x_{1,}x_2, ..., x_n$ data points, and the corresponding posterior $p(\theta|x_1, ..., x_n)$

- Now, a fresh additional data point x_{n+1} arrive.

- The posterior based on all available data is

$$p(\theta|x_{1,}..., x_{n+1}) \propto p(x_{n+1}|\theta, x_1, ..., x_n)p(\theta|x_1, ..., x_n).$$

- The following is thus equivalent:

- Analyzing the likelihood of all data $x_{1,}..., x_{n+1}$ with the prior based on no data $p(\theta)$

- Analyzing the likelihood of the fresh data point x_{n+1} with the 'prior' equal to the posterior based on the old data $p(\theta|x_1, ..., x_n)$.

• Yesterday's posterior is today's prior.

Conjugate priors

- Normal likelihood: Normal prior$\rightarrow$Normal posterior. (posterior belongs to the same distribution family as prior)

- Binomial likelihood: Beta prior$\rightarrow$Beta posterior.

- *Conjugate priors*: Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a class of sampling distributions. A family of distributions $\mathcal{P}$ is conjugate for $\mathcal{F}$ if
$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$
holds for all $p(x|\theta) \in \mathcal{F}$.

- *Natural conjugate prior*: $p(\theta) = c \cdot p(y_1, ..., y_n|\theta)$ for some constant c, i.e. the prior is of the same functional form as the likelihood.

- EXAMPLE (CONJUGATE PRIOR FOR POISSON MODEL). *Likelihood from iid Poisson sample* $y = (y_1, ..., y_n)$

$$p(y|\theta) = \left[\prod_{i=1}^{n} p(y_i|\theta)\right] \propto \theta^{(\sum_{i=1}^{n} y_i)} \exp(-\theta n),$$

so that the sum of counts $\sum_{i=1}^{n} y_i$ is a sufficient statistic for θ.
Natural conjugate prior for Poisson parameter θ

$$p(\theta) \propto \theta^{\alpha-1} \exp(-\theta\beta) \propto Gamma(\alpha, \beta)$$

which contains the info: $\alpha - 1$ counts in β observations.

Posterior for Poisson parameter θ. Multiplying the poisson likelihood and the Gamma prior gives the posterior

$$\begin{aligned} p(\theta|y_1, ..., y_n) &\propto \left[\prod_{i=1}^{n} p(y_i|\theta)\right] p(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} y_i} \exp(-\theta n)\theta^{\alpha-1} \exp(-\theta\beta) \\ &= \theta^{\alpha+\sum_{i=1}^{n} y_i-1} \exp[-\theta(\beta+n)], \end{aligned}$$

which is proportional to the $Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$ distribution. In summary

$$\begin{aligned} \text{Model: } & y_1, ..., y_n|\theta \overset{iid}{\sim} Po(\theta) \\ \text{Prior: } & \theta \sim Gamma(\alpha, \beta) \\ \text{Posterior: } & \theta|y_1, ..., y_n \sim Gamma(\alpha + \textstyle\sum_{i=1}^{n} y_i, \beta + n). \end{aligned}$$

Non-informative priors

- ... do not exist!

- ... may be improper and still lead to proper posterior

- Regularization priors

- Ideal communication. Present the posterior distributions for all possible priors.

- Practical communication - Reference priors.

Jeffreys' prior

- A common non-informative prior is Jeffreys' prior

$$p(\theta) = |I(\theta)|^{1/2},$$

where

$$J(\theta) = -E_{y|\theta}\left[\frac{d^2 \ln p(y|\theta)}{d\theta^2}\right]$$

is the expected Fisher information.

- EXAMPLE (JEFFREYS' PRIOR FOR BERNOULLI DATA):

$$y_1, ..., y_n|\theta \overset{iid}{\sim} Bern(\theta).$$

$$\ln p(y|\theta) = s \ln \theta + f \ln(1 - \theta)$$

$$\frac{d \ln p(y|\theta)}{d\theta} = \frac{s}{\theta} - \frac{f}{(1-\theta)}$$

$$\frac{d^2 \ln p(y|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

$$J(\theta) = \frac{E_{y|\theta}(s)}{\theta^2} + \frac{E_{y|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

$$p(\theta) = |J(\theta)|^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto Beta(\theta|1/2, 1/2).$$

Prediction

- We may use the estimated model for forecasting a future observation $\tilde{y}$.

- *Posterior predictive distribution* (y denotes available data at the time of forecasting)

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta$$

  where the last step holds if $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$.

- The uncertainty that comes from not knowing θ is represented in $p(\tilde{y}|y)$ by averaging over $p(\theta|y)$.

Gibbs sampling

- Easily implemented methods for sampling from multivariate distributions, $p(\theta_1, ..., \theta_k)$.

- Requirements: Easily sampled full conditional posteriors:
  - $p(\theta_1|\theta_2, \theta_3..., \theta_k)$
  - $p(\theta_2|\theta_1, \theta_3, ..., \theta_k)$
  - $\vdots$
  - $p(\theta_k|\theta_1, \theta_2, ..., \theta_{k-1})$

The Gibbs sampling algorithm

Step A: Choose initial values $\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_n^{(0)}$.

Step B: B_1 Draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_n^{(0)})$

B_2 Draw $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_n^{(0)})$

$\vdots$

B_n Draw $\theta_n^{(1)}$ from $p(\theta_n|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{n-1}^{(1)})$

Step C: Repeat Step B N times.

- The Gibbs draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are dependent, but arithmetic means converge to expected values

$$\frac{1}{N}\sum_{t=1}^{N} \theta_j^{(t)} \rightarrow E(x_j)$$
$$\frac{1}{N}\sum_{t=1}^{N} g(\theta^{(t)}) \rightarrow E[g(\theta)]$$

- More generally, the Gibbs sequence $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ converges in distribution to the target posterior $p(\theta_1, ..., \theta_k)$.

- $\theta_j^{(1)}, ..., \theta_j^{(N)}$ converge to the marginal distribution of θ_j, $p(\theta_j)$.

The Metropolis Algorithm

- Initialize with $\theta = \theta_0$

- For $t = 1, 2, ...$

  - Sample a proposal draw $\theta^* | \theta^{(t-1)} \sim J_t(\theta^*, \theta^{(t-1)})$

  - Accept θ^* with probability

$$r(\theta^*, \theta^{(t-1)}) = \text{min}\left[\frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}, 1\right].$$

    If the proposal is accepted, set $\theta^{(t)} = \theta^*$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$.

- We must be able to compute the posterior density $p(\theta|y)$ for any θ.

- The Metropolis algorithm works even if $p(\theta|y)$ is only known up to a proportionality constant as it simply cancels in $r(\theta^*, \theta^{(t-1)})$.

- The proposal, or jumping, distribution $J_t(\theta^*|\theta^{(t-1)})$ may vary from iteration to iteration.

- $J_t(\theta^*, \theta^{(t-1)})$ must be symmetric, i.e.

$$J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a) \text{ for all } \theta_a, \theta_b \text{ and } t.$$

- Every proposal that θ^* that lies uphill $(p(\theta^*|y) \geq p(\theta^{(t-1)}|y))$ is accepted with certainty. Downhill moves accepted with prob. $r(\theta^*, \theta^{(t-1)})$.

- Common choice of proposal distribution:

$$J_t(\theta^*|\theta^{(t-1)}) = N(\theta^{(t-1)}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = c^2 I^{-1}(\hat{\theta})$ and $I^{-1}(\hat{\theta})$ is the observed information matrix at the posterior mode (obtained either analytically or by numerical optimization prior to the posterior sampling). c is a tuning constant (see the 'optimal' value of c in Section 11.9).

The Linear Regression Model

- The ordinary linear regression model:

$$
\begin{aligned}
y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\
&\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).
\end{aligned}
$$

- Parameters $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$.

- Assumptions:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ (linear function)
  - $Var(y_i) = \sigma^2$ (homoscedasticity)
  - $Corr(y_i, y_j | X) = 0$, $i \neq j$.
  - Normality of ε_i.

- The linear regression model in matrix form

$$\underset{(n\times 1)}{y} = \underset{(n\times k)(k\times 1)}{X\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\begin{aligned} y &= \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\ X &= \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \end{aligned}$$

- Usually $x_{i1} = 1$, for all i. β_1 becomes the intercept.

- Likelihood:

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I_n)$$

- Standard non-informative prior: uniform on $(\beta, \log \sigma)$

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

- Joint posterior of β and σ^2:

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y).$$

- Conditional posterior of β :

$$\begin{aligned} \beta|\sigma^2, y &\sim N(\hat{\beta}, \sigma^2 V_\beta) \\ \hat{\beta} &= (X'X)^{-1}X'y \\ V_\beta &= (X'X)^{-1}. \end{aligned}$$

- Marginal posterior of σ^2 :

$$\begin{aligned} \sigma^2|y &\sim Inv\text{-}\chi^2(n-k, s^2) \\ s^2 &= \frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta}). \end{aligned}$$

- Marginal posterior of β :

$$\beta|y \sim t_{n-k}(\hat{\beta}, \sigma^2 V_\beta).$$

  which is proper if $n > k$ and X has full column rank.

- Simulate from the joint posterior by iteratively simulating from $p(\sigma^2|y)$ and $p(\beta|\sigma^2, y)$.

- Predictive distribution of response $\tilde{y}$ with known predictors $\tilde{X}$:

$$\tilde{y}|y, \tilde{X} = t_{n-k}[\tilde{X}\hat{\beta}, s^2(I + \tilde{X}V_\beta\tilde{X}')]$$

$$\begin{aligned} \text{Predictive Variance} &= s^2 I + \tilde{X} s^2 V_\beta \tilde{X}' \\ &= \varepsilon\text{-Variance} + \tilde{X}(\text{Posterior Variance of } \beta)\tilde{X}'. \end{aligned}$$