Bayesian Theory
Quick Notes

Ian S.W. Ma

Spring 2020

Contents

1 Statistics Basics **1**

2 Bayes Throrem **2**

3 Conjugate Priors **3**

4 Normal Distribution Priors **3**

5 Jeffrey's Prior **4**

6 Reference Prior **5**

7 Point Estimates **5**

8 Interval Estimates **6**

9 Classic Hypothesis Testing **6**

1 Statistics Basics

General Structure: Probability Space/Triple

Probability Space/Triple $(\Omega, \mathcal{F}, \mathbb{P})$

- Ω: **Sample Space**, set of all posible outcomes.
- $\mathcal{F}$: **Set of events** σ, each event is a subset of Ω
- $\mathbb{P}$ Probability of event $\sigma \in \mathcal{F}$

Conditional Probability

Conditional probability of A given B: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Leftrightarrow \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$

General properties

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

  Events are mutually exclusive $\Rightarrow \mathbb{P}(A \cap B) = 0$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$

  Events are independent $\Rightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$
- Law of Total Probability

  If events A_1, A_2, A_N are mutually exclusive and exhaustive ($\bigcup_{i=1}^{N} A_i = \Omega$) then:

$$\mathbb{P}(B) = \sum_{i=1}^{N} \mathbb{P}(A_i \cap B) = \sum_{i=1}^{N} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Random Variables

$X \sim$ Distribution	$\mathbb{E}[X]$	$\text{Var}[X]$
Normal(μ, σ^2)	μ	σ^2
Bernoulli(θ)	θ	$\theta(1-\theta)$
Binomial(θ)	$n\theta$	$n\theta(1-\theta)$
Poisson(μ)	μ	μ
Uniform(α, β)	$\frac{\alpha+\beta}{2}$	$\frac{(\alpha+\beta)^2}{12}$
Beta(α, β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma(α, β)	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Inv. Gamma(α, β)	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Random Vectors

$\mathbf{X} = X_1, ..., X_k \sim$ Distribution	$\mathbb{E}[X_i]$	$\text{Var}[X_i]$	$\text{Cov}[X_i, X_j]$
Multivariative Normal$(\boldsymbol{\mu}, \Sigma)$	μ_i	$\Sigma_{i,i}$	$\Sigma_{i,j}$
Multinomial$(n, \theta_1, ..., \theta_k)$	$n\theta_i$	$n(1-\theta_i)$	$-n\theta_i\theta_j$

2 Bayes Throrem

Discrete Case

$$
\begin{aligned}
\mathbb{P}(X = x_i|Y = y_j) &= \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} \\
&= \frac{\mathbb{P}(Y = y_j|X = x_i)\mathbb{P}(X = x_i)}{\mathbb{P}(Y = y_j)} \\
&= \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}} \\
&= \frac{\mathbb{P}(Y = y_j|X = x_i)\mathbb{P}(X = x_i)}{\sum_{k=1}^{n} \mathbb{P}(X = x_k, Y = y_j)} \\
&= \frac{\mathbb{P}(Y = y_j|X = x_i)\mathbb{P}(X = x_i)}{\sum_{k=1}^{n} \mathbb{P}(Y = y_j|X = x_k)P(X = x_k)} \\
\Rightarrow \mathbb{P}((X|Y)) &= \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\sum_x \mathbb{P}(X = x, Y)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}
\end{aligned}
$$

Continous Case

$$
f(X|Y) = \frac{f(X,Y)}{g(Y)} = \frac{g(Y|X)f(X)}{\int g(Y|X)f(X)dX}
$$

Bayesian Statistical Inference

- $L(\theta|y) = f(y|\theta)$: Likelihood
- $\pi(\theta)$: Prior
- $m(y) = p(y)$: Marginal distribution/Normalizing constant

$$
\text{Posterior Distribution: } p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int p(\theta, y)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta)
$$

Predictive Distrubutions

$$
\begin{aligned}
&\text{Prior Predictive Distribution:} && p(y^{new}) = \int_\theta f(y^{new}|\theta)\pi(\theta)d\theta \\
&\text{Posterior Predictive Distribution:} && p(y^{new}|y^{old}) = \int_\theta f(y^{new}|\theta, y^{old})p(\theta|y^{old})d\theta
\end{aligned}
$$

Bayes Theorem for Multiple Parameters $\Theta = \{\theta_1, ..., \theta_q\}$

$$
\mathbb{P}(\Theta|\boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{y}|\Theta)\mathbb{P}(\Theta)}{\mathbb{P}(\boldsymbol{y})} \propto \mathbb{P}(\boldsymbol{y}|\Theta)\mathbb{P}(\Theta)
$$

3 Conjugate Priors

Sample Distribution	Parameter	Prior
Bernoulli(θ)	θ	Beta(α, β)
Binomial(n, θ)	θ	Beta(α, β)
Poisson(θ)	θ	Gamma(α, β)
Exponential(θ)	θ	Gamma(α, β)
Normal(μ, σ_0^2)	μ	Normal(μ_h, σ_h^2)
Normal(μ_0, σ^2)	σ^2	Inverse Gamma(α, β)
Normal($\mu_0, 1/\tau$)	$\tau = 1/\sigma^2$	Gamma(α, β)

4 Normal Distribution Priors

Unknown μ, known σ^2

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu \sim \text{Normal}(\mu_0, \sigma_0^2) = \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) \Rightarrow m = \sigma^2/\sigma_0^2$
- Posterior:

$$\mu|\boldsymbol{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{\sigma^2}{m+n}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{\sigma^2}{m+n}\right)$$

$w = \frac{m}{m+n}$

Unknown μ, known τ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) = \text{Normal}\left(\mu_0, \frac{1}{\tau m}\right)$
- Posterior:

$$\mu|\boldsymbol{y}, \sigma^2 \sim \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{1}{\tau(m+n)}\right) = \text{Normal}\left((1-w)\mu_0 + w\bar{y}, \frac{1}{\tau(m+n)}\right)$$

$w = \frac{m}{m+n}$

Unknown τ, known μ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\tau \sim \text{Gamma}(\alpha, \beta)$
- Posterior:

$$\tau|\boldsymbol{y}, \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{z^2}{2}\right)$$

$z^2 = \sum_i (y_i - \mu)^2$

Unknown σ^2, known μ

- Sample $\boldsymbol{y} \in \mathbb{R}^n$
- Prior $\mu, \sigma^2 \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{\kappa}\right) \times \text{Inverse Gamma}(\alpha, \beta)$
- Posterior:

  $$\mu, \sigma^2|\boldsymbol{y} \sim \text{Normal}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{\sigma^2}{\kappa + n}\right)\times\text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa n(\bar{y} - \mu_0)^2}{2(\kappa + n)}\right)$$

  $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$
  $\kappa = \sigma^2(\alpha - 1)/\beta$ or given (seriously it's not written anywhere Ken, I'm just guessing here man)

5 Jeffrey's Prior

Definition

Jeffery Prior: $\pi_{JP}(\theta) \propto \sqrt{I(\theta|y)}$
Fisher Information:

$$I(\theta|y) = \mathbb{E}\left[\left(\frac{d \log f(y|\theta)}{d\theta}\right)^2\right] = -\mathbb{E}\left[\frac{d^2 \log f(y|\theta)}{d\theta^2}\right]$$

1:1 transformation of Jeffrey's Prior

$$\pi(\phi) = \pi_{JP}(\theta)\left|\frac{d\theta}{d\phi}\right|$$

Jeffrey's Prior for Multivariate Parameter Vector

Gradient of log likelihood:

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \partial_{\theta_1} \log f(x) \\ \vdots \\ \partial_{\theta_n} \log f(x) \end{bmatrix}$$

Hessian Matrix of log likelihood:

$$H(\boldsymbol{\theta}) = \text{Jacobian}\left[S(\boldsymbol{\theta})\right] = \begin{bmatrix} \partial_{\theta_1}\partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_1}\partial_{\theta_n} \ln f(x) \\ \vdots & \ddots & \vdots \\ \partial_{\theta_n}\partial_{\theta_1} \ln f(x) & \cdots & \partial_{\theta_n}\partial_{\theta_n} \ln f(x) \end{bmatrix}$$

Fisher Information: $I(\theta|x) = -\mathbb{E}\left[H(\theta)\right]$
Jeffrey's Prior: $\pi_{JP} \propto \sqrt{\det(I(\theta|x))}$

6 Reference Prior

Kullback-Leibler (KL) divergence

A measure of the difference between two pmfs/pdfs. When $KL(f, g) = 0$, the two distributions are idetical.

Properties

- $KL(f, g) \neq KL(g, f)$
- $KL(fmg) \geq 0$

Continous parameter x

$$KL(f, g) = \int \ln \left[\frac{f(x)}{g(x)}\right] dx = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

Discrete parameter x

$$KL(f, g) = \sum_{x \in X} \ln \left[\frac{f(x)}{g(x)}\right] = \mathbb{E}_f[\log f(X)] - \mathbb{E}_g[\log f(X)]$$

Reference Prior

Read week 4 notes (2.2)

7 Point Estimates

Components of Decision Theory with emphasis on parameter estimation

- State Spcae Θ, unknown true value $\theta \in \Theta$
- Action Space $\mathcal{A} \ni a$, sometimes $\mathcal{A} = \Theta$
- Sampling Distribution $f(\boldsymbol{y}|\theta)$
- Loss Function $\mathcal{L}(a|\theta)$
- Risk $R_\theta(a|\boldsymbol{y}) = \mathbb{E}_\theta[L(a|\theta, y)] = \int_{\theta \in \Theta} \mathcal{L}(a|\theta)p(\theta|\boldsymbol{y})d\theta$
- Bayes Estimator of a parameter $\hat{\theta}_{BE} = \arg\min_{\hat{\theta} \in \Theta} R_\theta(\hat{\theta}|\boldsymbol{y})$

Common Loss Functions and Corresponding Estimators

- **Squared Error Loss**: $\mathcal{L}(\theta|\hat{\theta}) = (\theta - \hat{\theta})^2$, Bayes Estimator $\hat{\theta} = \mathbb{E}[\theta|\boldsymbol{y}]$
- **Absolute Error Loss**: $\mathcal{L}(\theta|\hat{\theta}) = |\theta - \hat{\theta}|$, Bayes Estimator $\hat{\theta} = \theta_{0.5}$, $(\mathbb{P}(\theta \leq \theta_{0.5}) = 0.5)$
- **0-1 Loss**: $\mathcal{L}(\theta|\hat{\theta}) = I(\theta \neq \hat{\theta})$, Bayes Estimator $\hat{\theta}$ is the posterior mode

8 Interval Estimates

$$P \times 100\% \text{ Bayesian Credible interval} = [LB, UB] \text{ where } \int_{LB}^{UB} p(\theta, \boldsymbol{y})d\theta = P$$

If loooking for one side-confidence bounds then $LB = -\infty$ or $UB = \infty$

Symmetric Credicble Interval

$$[LB, UB] \text{ where } \mathbb{P}(\theta \leq LB) = \mathbb{P}(\theta \geq UB) = \alpha/2 = (1-P)/2$$

Highest Posterior Density Interval (HPDI)

Credible Interval where all values outside the interval has a density smaller than any of the values in the interval.

Credible Regions

$$\iint p(\theta_1, \theta_2|\boldsymbol{y})d\theta_1 d\theta_2 = 1 - \alpha = P$$

9 Classic Hypothesis Testing

1. Assume hypothesis H_0 is true
2. Calculate test statistic $T(\boldsymbol{y}_{obs})$ based on observed sample data regarding to H_0 and H_1
3. Conditional on H_0 being true, p-value $= \mathbb{P}(T(\boldsymbol{y})$ more extreme than $T(\boldsymbol{y}_{obs})|\theta, H_0)$
4. Reject H_0 and accept H_1 for sufficiently small p-values, do not reject H_0 otherwise