MAT 4381

Bayesian Inference

Study Guide

Winter 2024

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayesian vs. Frequentist

Frequentist: The parameter is not a random variable

Bayesian: The parameter is a random variable

Bayes Theorem

Ω be the sample space

$E_1, E_2, \ldots$ sequence of exhaustive events

$A \subset \Omega$ another event

$$P(E_k|A) = \frac{P(A|E_k)P(E_k)}{\sum_k P(A|E_k)P(E_k)}$$

Bernoulli Trials

X takes value 1 if event occurs and 0 if event do not occur

Bernoulli Distribution: $Ber(\theta)$

$P(X=1)=\theta$

$P(X=0)=1-\theta$

$P(X=x)=f(x|\theta)=\theta^x(1-\theta)^{1-x} I(x\in\{0,1\})$

$E[X]=\sum_{x=0}^{1} x f(x|\theta) = \theta$

$Var(X)=\theta(1-\theta)$

Binomial Distribution

$X_1,\ldots,X_n \overset{iid}{\sim} Ber(\theta)$

$Y=\sum_{i=1}^{n} X_i \in\{0,1,2,\ldots,n\} \sim Bin(n,\theta)$

$P(Y=y)=\binom{n}{y}\theta^y(1-\theta)^{n-y} I(y\in\{0,1,\ldots,n\})$

$E[Y]=n\theta \quad Var(Y)=n\theta(1-\theta)$

Bayes Formula for Distributions

$$f(\theta|y)=\frac{f(y|\theta)g(\theta)}{f_1(y)}$$

If θ is discrete with p.m.f. $g(\theta)$

$$f(\theta|y)=\frac{f(y|\theta)g(\theta)}{f_1(y)}=\frac{f(y|\theta)g(\theta)}{\sum_j f(y|\theta_j)g(\theta_j)}$$

If $\theta \sim g(\theta)$ is continuous

$$f(\theta|y)=\frac{f(y|\theta)g(\theta)}{f_1(y)}=\frac{f(y|\theta)g(\theta)}{\int f(y|\theta)g(\theta)d\theta}$$

The denominator is sometimes called the proportionality constant

$f(\theta|y)=Cf(y|\theta)g(\theta)$

or $f(\theta|y)\propto f(y|\theta)g(\theta)$ = likelihood × prior

C some constant and $\propto$ means "proportional to"

$g(\theta)$: prior distribution for θ

$f(\theta|y)$: posterior distribution

Conjugate Prior

- A conjugate prior is also a convenient prior, since it allows one to easily calculate the posterior distribution

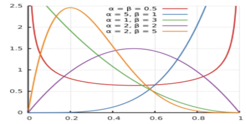
Beta Distribution

$$g(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} I(0<\theta<1)$$

The distribution is the $Beta(\alpha,\beta)$ distribution

We use the fact that

$$\frac{\Gamma(\alpha)}{\beta^\alpha}=\int_0^\infty x^{\alpha-1}e^{-\beta x}dx,\ \alpha,\beta>0$$



The Beta distribution is a conjugate prior for binomial distribution

If $(Y|\Theta=\theta)\sim f(y|\theta)$

$f(y|\theta)=\binom{n}{y}\theta^y(1-\theta)^{n-y} I(y\in\{0,1,\ldots,n\})$

and $(\theta|\alpha,\beta)\sim Beta(\alpha,\beta)$

we have

$f(\theta|y) \propto f(y|\theta)g(\theta)$

In other words,

$f(\theta|y)\propto\theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

The posterior is:

$$f(\theta|y)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

If the prior is $Beta(\alpha,\beta)$, then the posterior is $Beta(\alpha+y,\beta+n-y)$

If $\Theta\sim Beta(\alpha,\beta)$, then

$$E[\Theta]=\frac{\alpha}{\alpha+\beta}$$

$$Var(\Theta)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\text{Mode: } \hat{\theta}=\frac{\alpha-1}{\alpha+\beta-2}$$

i) For large values of α and β mode and mean are roughly the same.

ii) Median for $Beta(\alpha,\beta)$ cannot be written in a closed form

iii) Peak of mode of beta density gets sharper as $\alpha+\beta$ increases

iv) We need to use numerical methods to calculate the med for different values of α and β

Credible Region

- We want to find I s.t. $P(\Theta\in I)=1-\alpha$, where $1-\alpha$ is the confidence level
- We can also use simulation method

Choosing Beta function

- First, consider priors that "don't express a strong" opinion.
- Later, consider priors which incorporate a strong prior belief. If there are no preferences on the value for θ, then one might consider using a "flat prior". That is, the uniform distribution between 0 and 1 ($Beta(1,1)$)
- In some sense, this implies that each value of θ as equally likely
- Since the prior is flat, the posterior distribution has the same shape as the likelihood multiplied by a constant C

$$h(\theta|x)=Cf(x|\theta)g(\theta)\propto f(x|\theta)$$

- Since the uniform distribution is $Beta(1,1)$, then the posterior is $Beta(y+1, n-y+1)$

Making Bayesian Inference

The posterior contains the results of the analysis

To communicate the result, we can either present the entire posterior distribution, or give summary statistics of the posterior

For credible regions, there are 2 common types:

- Smallest interval
- Equal tail area

Jeffreys Prior

Jeffreys proposed that an acceptable "non-informative prior finding principle" should be invariant under monotone transformations of the parameter

The Jeffrey's prior satisfying this invariance is proportional to $\sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information on θ. Can be generalized when θ is k-dimensional to $\sqrt{det(I(\theta))}$

$$I(\theta)=-E\left[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right]$$

For k-dimensional, we have

$$I(\theta)=E\left[\left(\frac{\partial^2 \ln f(x|\theta)}{\partial\theta_i\theta_j}\right)_{(i,j)}\right]$$

Multinomial Distribution

$(X_1,\ldots,X_k)\sim Multinomial(n,p_1,\ldots,p_k)$

$p_k=1-p_1-p_2-\cdots-p_{k-1}$

Models probability of counts for each side of an experiment with k exhaustive outcomes

$$f(x_1,x_2,\ldots,x_k)=\binom{n}{x_1,x_2,\ldots,x_{k-1}}p_1^{x_1}p_2^{x_2}\cdots p_{k-1}^{x_{k-1}}(1-p_1-p_2-\ldots-p_{k-1})^{x_k}$$

where $x,y=0,1,2,\ldots,n,\ x_1+x_2+\ldots+x_{k-1}\le n$

Simple Case: Trinomial Distribution

$(X, Y)\sim Multinomial(n,p_1,p_2)$

$f(x,y)=\binom{n}{x,y}p_1^x p_2^y(1-p_1-p_2)^{n-x-y}$

$x,y=0,1,2,\ldots,n,\ x+y\le n$

We can also get that

$X\sim Bin(n,p_1)$

$Y\sim Bin(n,p_2)$

$Y|X=x \sim Bin\left(n-x,\frac{p_2}{1-p_1}\right)$

$E[Y|X=x]=(n-x)\frac{p_2}{1-p_1}$

$Var[Y|X=x]=(n-x)\left(\frac{p_2}{1-p_1}\right)\left(1-\frac{p_2}{1-p_1}\right)$

$E[XY]=\frac{p_2}{1-p_1}E[X(n-X)]=-np_1p_2$

$Cov(X,Y)=-n(n+1)p_1p_2$

$$\rho(X,Y)=\frac{-n(n+1)p_1p_2}{\sqrt{np_1(1-p_1)np_2(1-p_2)}}=\frac{-(n+1)p_1p_2}{\sqrt{p_1p_2(1-p_1)(1-p_2)}}$$

Jeffreys Prior for Trinomial Distribution

$$g(p_1,p_2)\propto\frac{1}{\sqrt{p_1p_2(1-p_1-p_2)}}$$

Special case of Dirichlet distribution $Dir(1/2,1/2,1/2)$

Dirichlet Distribution

$(X_1,X_2,\ldots,X_{k-1})\sim Dir(\alpha_1,\alpha_2,\ldots,\alpha_k)$

$$f(x_1,\ldots,x_{k-1})=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\left(\prod_{i=1}^{k-1}x_i^{\alpha_i-1}\right)\left(1-\sum_{i=1}^{k-1}x_i\right)^{\alpha_k-1}$$

Defined on $S=\{(x_1,\ldots,x_{k-1}): \sum_{i=1}^{k-1}x_i<1,\ x_i>0,\ i=1,\ldots,k\}$

Posterior for Trinomial

$(\theta_1,\ldots,\theta_{k-1}|X_1=x_1,\ldots,X_{k-1}=x_{k-1})\sim Dir(\alpha_1+x_1,\ldots,\alpha_{k-1}+x_{k-1},n-\sum x_i+\alpha_k)$

$$E[\Theta|X_1=x_1,\ldots,X_{k-1}=x_{k-1}]=\left[\frac{\alpha_1+x_1}{\sum\alpha_i+n},\ldots,\frac{\alpha_k+x_k}{\sum\alpha_i+n}\right]$$

$\Theta=[\theta_1\ \theta_2\ \cdots\ \theta_{k-1}\ \theta_k]^T$

$\theta_k=1-\theta_1-\ldots-\theta_{k-1} \quad x_k=n-x_1-\cdots-x_{k-1}$

Notice: $(\theta_1,\theta_2)\sim Dir(\alpha_1,\alpha_2,\alpha_3+\ldots+\alpha_k)$

Basic Properties of Dirichlet Distribution

$(Y_1,\ldots,Y_{k-1})\sim Dir(\alpha_1,\ldots,\alpha_k) \quad \alpha_0=\sum_{i=1}^{k}\alpha_i$

$$E[Y_i]=\frac{\alpha_i}{\alpha_0} \quad Var[Y_i]=\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$$

$$Cov(X_i,X_j)=-\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$$

Conjugate Prior and Posterior for Poisson

$$X_1,\ldots,X_n \overset{iid}{\sim} f(x|\theta)=\frac{\exp(-\theta)\theta^x}{x!}I(x=0,1,\ldots)$$

$$\text{Likelihood: } \prod_{i=1}^{n}f(x_i|\theta)=\prod_{i=1}^{n}\frac{\exp(-\theta)\theta^{x_i}}{x!}\propto\exp(-n\theta)\theta^y$$

$y=\sum_{i=1}^{n}x_i$

$$\text{Conjugate Prior: } g(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)I(\theta>0)$$

$$\text{Posterior: } h(\theta|y)\propto\theta^{y+\alpha+1}\exp(-(\beta+n)\theta)I(\theta>0)$$

$$h(\theta|y)=\frac{(\beta+n)^{y+\alpha}}{\Gamma(y+\alpha)}\theta^{y+\alpha+1}\exp(-(\beta+n)\theta)I(\theta>0)$$

$$E[\Theta|Y=y]=\frac{y+\alpha}{n+\beta}$$

$$Var[\Theta|Y=y]=\frac{y+\alpha}{(n+\beta)^2}$$

$$Mode(\Theta|Y=y)=\frac{y+\alpha-1}{n+\beta}$$

Jeffrey's Prior for Poisson

$$f(x|\lambda) = \frac{\exp(-n\lambda)\,\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$$-E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] = \frac{n}{\lambda}$$

Jeffrey's prior: $g(\lambda) \propto \frac{1}{\sqrt{\lambda}}$

This leads to an improper prior since we have

$$\int_0^\infty \frac{d\lambda}{\sqrt{\lambda}} = \infty$$

No normalizing constant will turn it into a prior p.d.f.

We can still use it to calculate the posterior

$$\lambda|X \propto \exp(-n\lambda)\,\lambda^{y+1/2-1}$$

The use of improper prior leads to Gamma posterior

$$\lambda|X \sim \frac{1}{\Gamma(\sum_{i=1}^{n} x_i + 1/2)}\lambda^{\sum_{i=1}^{n} x_i + \frac{1}{2} - 1}\exp(-n\lambda)$$

$$E[\lambda|x_1,\dots,x_n] = \frac{\sum_{i=1}^{n} x_i + 0.5}{n}$$

$$Var[\lambda|x_1,\dots,x_n] = \frac{\sum_{i=1}^{n} x_i + 0.5}{n^2}$$

Normal Distribution

$(Y_1,\dots,Y_n|\mu) \sim f(y_1,\dots,y_n|\mu) \propto \exp\left(-\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\right)$

σ^2 known

$$f(y_1,\dots,y_n|\mu) \propto \exp\left(-\frac{\mu^2-2\mu\bar{y}}{2\sigma^2/n}\right)$$

The Jeffrey's prior is

$$g(\mu) \propto 1$$

This constitutes an improper prior

However, we can still calculate the posterior

$$(\mu|Y_1=y_1,\dots,Y_n=y_n) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

The normal prior

$Y_1,\dots,Y_n \overset{iid}{\sim} N(\theta,\sigma^2)$

σ^2 known

Prior: $g(\theta) \propto \exp(-(\theta-\mu)^2/\tau^2)$

Likelihood: $\prod_{i=1}^{n} f(y_1,\dots,y_n|\mu) \propto f(\bar{y}|\mu)$
$\propto \exp\left(-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}\right)$

Posterior for θ

$$\begin{aligned}
h(\theta|y_1,y_2,\dots,y_n) &\propto \exp\left(-(\theta-\mu)^2/\tau^2\right)\cdot\exp\left(-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}\right)\\
&\propto \exp\left(\frac{(\bar{y}-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu)^2}{2\tau^2}\right)\\
&\propto \exp\left(-\frac{\theta^2-2\theta\bar{y}}{2\sigma^2/n} - \frac{\theta^2-2\mu\theta}{2\tau^2}\right)\\
&\propto \exp\left(-\frac{\theta^2(\tau^2+\sigma^2/n)-2\theta(\tau^2\bar{y}+\mu\sigma^2/n)}{2\tau^2\sigma^2/n}\right)\\
&\propto \exp\left(-\frac{\theta^2-2\theta(\tau^2\bar{y}+\mu\sigma^2/n)/(\tau^2+\sigma^2/n)}{\frac{2\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}}\right)
\end{aligned}$$

$$(\theta|y_1,\dots,y_n) \sim N\left(\frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+\sigma^2/n}, \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}\right)$$

$$E[\theta|Y=y] = \frac{\tau^2\bar{y}+\mu\sigma^2/n}{\tau^2+\sigma^2/n}$$

$$MAP = E[\theta|y_1,\dots,y_n] = \text{Median}(\theta|Y=y)$$

Posterior precision: $\frac{\tau^2+\sigma^2/n}{\tau^2\sigma^2/n} = \frac{1}{\sigma^2/n} + \frac{1}{\tau^2}$

Posterior precision is equal to prior precision plus observation precision

Since $E[\theta|y_1,\dots,y_n] = \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu + \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{y}$, the posterior mean is the weighted average of the prior mean and the mean of observations, where the weights are the proportions of the precision to the posterior precision.