The Bayesian Approach to Statistical Modeling

Version 1.5: 2003/08/01 13:27:47.817 GMT-5

Rob Nowak
Clayton Scott

Abstract

The Bayesian Approach to Statistical Modeling

Example 1:

$x_n = A + W_n \forall n, n = \{1, \dots, N\}$

Prior distribution allows us to incorporate prior information regarding unknown paremter– probable values of parameter are supported by prior. Basically, the prior reflects what we believe "Nature" will probably throw at us.

1 Elements of Bayesian Analysis

(a) joint distribution

$$p(\mathbf{x}, \theta) = p(\mathbf{x} \mid \theta)\, p(\theta)$$

(b) marginal distributions

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \theta)\, p(\theta)\, d\theta$$

$$p(\theta) = \int p(\mathbf{x} \mid \theta)\, p(\theta)\, d\mathbf{x}$$

where $p(\theta)$ is a **prior**.

[redacted]

Figure 1

[*] http://creativecommons.org/licenses/by/1.0

[redacted]

Figure 2

[redacted]

Figure 3: This reflects prior knowledge that most probable values of θ are close to $\frac{\alpha}{\alpha+\beta}$.

(c) posterior distribution

$$p\left(\theta\ |\mathbf{x}\right)=\frac{p\left(\mathbf{x},\theta\right)}{p\left(\mathbf{x}\right)}=\frac{p\left(\mathbf{x}\ |\theta\right)p\left(\theta\right)}{\int p\left(\mathbf{x}\ |\theta\right)p\left(\theta\right)d\mathbf{x}}$$

Example 2:

$$p\left(\mathbf{x}\ |\theta\right)=\left(\begin{array}{c}n\\ x\end{array}\right)\theta^{x}(1-\theta)^{n-x}\forall\theta,\theta\in[0,1]$$

which is the Binomial likelihood.

$$p\left(\theta\right)=\frac{1}{B\left(\alpha,\beta\right)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

which is the Beta prior distriubtion and where $B\left(\alpha,\beta\right)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

joint density:

$$p\left(\mathbf{x},\theta\right)=\frac{\left(\begin{array}{c}n\\ x\end{array}\right)}{B\left(\alpha,\beta\right)}\theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1}$$

marginal density:

$$p\left(\mathbf{x}\right)=\left(\begin{array}{c}n\\ x\end{array}\right)\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\frac{\Gamma\left(\alpha+x\right)\Gamma\left(n-x+\beta\right)}{\Gamma\left(\alpha+\beta+n\right)}$$

posterior density:

$$p\left(\theta\ |\mathbf{x}\right)=\frac{\theta^{\alpha+x-1}\theta^{\beta+n-x-1}}{B\left(\alpha+x,\beta+n-x\right)}$$

where $B\left(\alpha+x,\beta+n-x\right)$ is the Beta density with parameters $\alpha'=\alpha+x$ and $\beta'=\beta+n-x$

2 Bayesian Estimation

We are interested in estimating θ given the observation $\mathbf{x}$. Naturally then, any estimation strategy will be based on the posterior distribution $p\left(\theta\ |\mathbf{x}\right)$. Furthermore, we need a criterion for assessing the quality of potential estimators.

3 Loss

The quality of an estimate $\hat{\theta}$ is measured by a real-valued **loss function**: $L\left(\theta,\hat{\theta}\right)$. For example, squared error or quadratic loss is simply $L\left(\theta,\hat{\theta}\right)=\left(\theta-\hat{\theta}\right)^T\left(\theta-\hat{\theta}\right)$

4 Expected Loss

Posterior Expected Loss:

$$E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]=\int L\left(\theta,\hat{\theta}\right)p\left(\theta\ |\mathbf{x}\right)d\theta$$

Bayes Risk:

$$\begin{array}{rcl} E\left[L\left(\theta,\hat{\theta}\right)\right] & = & \int\int L\left(\theta,\hat{\theta}\right)p\left(\theta\ |\mathbf{x}\right)p\left(\mathbf{x}\right)d\theta d\mathbf{x}\\ & = & \int\int L\left(\theta,\hat{\theta}\right)p\left(\mathbf{x}\ |\theta\right)p\left(\theta\right)d\mathbf{x}d\theta\\ & = & E\left[E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]\right]\end{array}\quad(1)$$

The "best" or optimal estimator given the data $\mathbf{x}$ and under a specified loss is given by

$$\hat{\theta}=\underset{\theta}{argmin}E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]$$

Example 3: Bayes MSE

$$\text{BMSE}\left(\hat{\theta}\right)\equiv\int\int\left(\theta-\hat{\theta}\right)^2p\left(\theta\ |\mathbf{x}\right)d\theta p\left(\mathbf{x}\right)d\mathbf{x}$$

Since $p\left(\mathbf{x}\right)\geq 0$ for every $\mathbf{x}$, minimizing the inner integral $\int\left(\theta-E\left[\theta\right]\right)^2p\left(\theta\ |\mathbf{x}\right)d\theta=E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]$ (where $E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]$ is the posterior expected loss) for each $\mathbf{x}$, minimizes the overall BMSE.

$$\begin{array}{rcl}\frac{\partial}{\partial\hat{\theta}}\left(\int\left(\theta-\hat{\theta}\right)^2p\left(\theta\ |\mathbf{x}\right)d\theta\right) & = & \int\frac{\partial}{\partial\hat{\theta}}\left(\left(\theta-\hat{\theta}\right)^2p\left(\theta\ |\mathbf{x}\right)\right)d\theta\\ & = & -2\int\left(\theta-\hat{\theta}\right)p\left(\theta\ |\mathbf{x}\right)d\theta\end{array}\quad(2)$$

Equating this to zero produces

$$\hat{\theta}=\int\theta p\left(\theta\ |\mathbf{x}\right)d\theta\equiv E\left[\theta\mid\ \mathbf{x}\right]$$

The conditional (also called **posterior**) mean of θ given $\mathbf{x}$!

Example 4:

$$x_n=A+W_n\forall n,n=\{1,\dots,N\}$$
$$W_n\sim\mathcal{N}\left(0,\sigma^2\right)$$

prior for unknown parameter A:

$$p\left(a\right)=U\left(-A_0,A_0\right)$$

$$p\left(\mathbf{x}\mid A\right)=\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(\left(x_n-A\right)^2\right)}$$

$$p\left(A\mid\mathbf{x}\right)=\begin{cases}\dfrac{\frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(\left(x_n-A\right)^2\right)}}{\int_{-A_0}^{A_0}\frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(\left(x_n-A\right)^2\right)}dA}\text{ if }|A|\leq A_0\\ 0\text{ if }|A|>A_0\end{cases}$$

Minimum Bayes MSE Estimator:

$$\begin{aligned}\hat{A}&=E\left[A\mid\mathbf{x}\right]\\&=\int_{-\infty}^{\infty}ap\left(A\mid\mathbf{x}\right)dA\\&=\frac{\int_{-A_0}^{A_0}A\frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(\left(x_n-A\right)^2\right)}dA}{\int_{-A_0}^{A_0}\frac{1}{2A_0\left(2\pi\sigma^2\right)^{\frac{N}{2}}}e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}\left(\left(x_n-A\right)^2\right)}dA}\end{aligned}\tag{3}$$

Notes

1. No closed-form estimator
2. As $A_0\rightarrow\infty$, $\hat{A}\rightarrow\frac{1}{N}\sum_{n=1}^{N}x_n$
3. For smaller A_0, truncated integral produces an $\hat{A}$ that is a funciton of $\mathbf{x}$, σ^2, and A_0
4. As N increases, $\frac{\sigma^2}{N}$ decreases and posterior $p\left(A\mid\mathbf{x}\right)$ becomes tightly clustered about $\frac{1}{N}\sum x_n$. This implies $\hat{A}\rightarrow\frac{1}{N}\sum x_n$ as $n\rightarrow\infty$ (the data "swamps out" the prior)

5 Other Common Loss Functions

5.1 Absolute Error Loss

(Laplace, 1773)

$$L\left(\theta,\hat{\theta}\right)=|\theta-\hat{\theta}|$$

$$\begin{aligned}E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]&=\int_{-\infty}^{\infty}|\theta-\hat{\theta}|p\left(\theta\mid\mathbf{x}\right)d\theta\\&=\int_{-\infty}^{\hat{\theta}}\left(\hat{\theta}-\theta\right)p\left(\theta\mid\mathbf{x}\right)d\theta+\int_{\hat{\theta}}^{\infty}\left(\theta-\hat{\theta}\right)p\left(\theta\mid\mathbf{x}\right)d\theta\end{aligned}\tag{4}$$

Using integration-by-parts if can be shown that

$$\int_{-\infty}^{\hat{\theta}}\left(\hat{\theta}-\theta\right)p\left(\theta\mid\mathbf{x}\right)d\theta=\int_{-\infty}^{\hat{\theta}}P\left(\theta<y\mid\mathbf{x}\right)dy$$

where $P\left(\theta<y\mid\mathbf{x}\right)$ is a cumulative distribution.

$$\int_{\hat{\theta}}^{\infty}\left(\theta-\hat{\theta}\right)p\left(\theta\mid\mathbf{x}\right)d\theta=\int_{\hat{\theta}}^{\infty}P\left(\theta>y\mid\mathbf{x}\right)dy$$

So,

$$E\left[L\left(\theta,\hat{\theta}\right)\mid\mathbf{x}\right]=\int_{-\infty}^{\hat{\theta}}P\left(\theta<y\mid\mathbf{x}\right)dy+\int_{\hat{\theta}}^{\infty}P\left(\theta>y\mid\mathbf{x}\right)dy$$

Take the derivative with respect to $\hat{\theta}$ implies $P\left(\theta<\hat{\theta}\mid\mathbf{x}\right)=P\left(\theta>\hat{\theta}\mid\mathbf{x}\right)$ which implies that the optimal $\hat{\theta}$ under absolute error loss is **posterior median**.

[redacted]

Figure 4

5.2 '0-1' Loss

$$L\left(\theta, \hat{\theta}\right) = \begin{cases} 0 \text{ if } \hat{\theta} = \theta \\ 1 \text{ if } \hat{\theta} \neq \theta \end{cases} = I_{left\{\hat{\theta} \neq \theta right\}}$$

$$E\left[L\left(\theta, \hat{\theta}\right) \mid \mathbf{x}\right] = E\left[\mid \mathbf{x}\right] = \Pr\left[\hat{\theta} \neq \theta \ \mid \mathbf{x}\right]$$

which is the probability that $\hat{\theta} \neq \theta$ given $\mathbf{x}$. To minimize '0-1' loss we must choose $\hat{\theta}$ to be the value of θ with the highest posterior probability, which implies $\hat{\theta} \neq \theta$ with the smallest probability. The optimal estimator $\hat{\theta}$ under '0-1' loss is the **maximum *a posteriori*** (MAP) estimator–the value of θ where $p\left(\theta \ \mid \mathbf{x}\right)$ is maximized.