Estadística Bayesiana

Irving Gómez Méndez

UNIVERSIDAD IBEROAMERICANA
CIUDAD DE MÉXICO ®

Aproximación Normal

Aproximación Normal

Sea $Y_1, \ldots, Y_n \overset{iid}{\sim} f(Y)$, pero que nosotros modelamos como $p(Y|\theta)$. Además, $p(\theta)$ es la distribución previa de nuestro modelo. Entonces

$$p(\mathbf{Y}|\theta) = \prod_{i=1}^{n} p(Y_i|\theta)$$

será la verosimilitud de la muestra observada.
Sea θ_0 el valor que minimiza la divergencia de Kullback-Leibler entre $f(Y)$ y $p(Y|\theta)$,

$$\begin{aligned} KL(\theta) &= \mathbb{E}_{Y\sim f}\left[\log\left(\frac{f(Y)}{p(Y|\theta)}\right)\right] \\ &= \int_{\mathcal{Y}} \log\left(\frac{f(Y)}{p(Y|\theta)}\right) f(Y)dY \end{aligned}$$

Vamos a demostrar que, cuando n aumenta, la distribución posterior $p(\theta|\mathbf{Y})$ se concentra alrededor de θ_0. Para ello, primero consideraremos el caso en que Θ es un espacio discreto.

Teorema

Si el espacio parametral Θ es finito y $\mathbb{P}(\theta = \theta_0) > 0$, entonces $\mathbb{P}(\theta = \theta_0|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

Demostración

Considere el logaritmo del cociente de posteriores:

$$\log\left(\frac{p(\theta|\mathbf{Y})}{p(\theta_0|\mathbf{Y})}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

Note que

$$\log\left(\frac{p(\theta)}{p(\theta_0)}\right)$$

es una constante, al no depender de n.
Por otro lado, note que

$$\begin{aligned}
&\mathbb{E}_{Y\sim f}\left[\log\left(\frac{p(Y|\theta)}{p(Y|\theta_0)}\right)\right] \\
=&\mathbb{E}\left[\log f(Y) - \log p(Y|\theta_0) - \log f(Y) + \log p(Y|\theta)\right] \\
=&KL(\theta_0) - KL(\theta)
\end{aligned}$$

y por ley fuerte de grandes números se cumple que

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right) \xrightarrow[n\to\infty]{c.s.} KL(\theta_0) - KL(\theta) < 0$$

Por lo tanto,

$$\sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right) \xrightarrow[n\to\infty]{} -\infty,$$

luego

$$\frac{p(\theta|\mathbf{Y})}{p(\theta_0|\mathbf{Y})} \xrightarrow[n\to\infty]{} 0$$

y

$$p(\theta|\mathbf{Y}) \xrightarrow[n\to\infty]{} 0, \quad \text{para todo } \theta \neq \theta_0$$

Como la suma de las probabilidades tiene que ser 1, concluimos que

$$p(\theta_0|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$$

$\blacksquare$

Teorema

Sea Θ un espacio compacto y A un vecindario de θ_0 tal que $\mathbb{P}(\theta \in A) > 0$, entonces $\mathbb{P}(\theta \in A|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

Demostración

Como Θ es compacto, entonces existe una cobertura finita de Θ y se puede construir de tal manera que A es el único vecindario que incluye a θ_0. Usando el teorema anterior se puede demostrar que la probabilidad posterior para cualquier vecindario que no sea A tiende a 0 cuando $n \to \infty$ y $\mathbb{P}(\theta \in A|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

$\blacksquare$

Teorema

Bajo condiciones de regularidad (que incluyen que θ_0 no esté en la frontera de Θ), la distribución posterior de θ es aproximadamente normal con media θ_0 y varianza $(nJ(\theta_0))^{-1}$

Demostración

Sea $\hat{\theta}$ la moda de la distribución posterior. Luego,

$$\log p(\theta|\mathbf{Y}) = \log p(\hat{\theta}|\mathbf{Y}) + \frac{1}{2}(\theta - \hat{\theta})^2 \left. \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{Y}) \right|_{\theta=\hat{\theta}}$$

Note que

$$\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\bigg|_{\theta=\hat{\theta}} = \frac{d^2}{d\theta^2}\log p(\theta)\bigg|_{\theta=\hat{\theta}} + \sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(y_i|\theta)\bigg|_{\theta=\hat{\theta}}$$

Por ley fuerte de grandes números y los teoremas anteriores, tenemos que

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(y_i|\theta)\bigg|_{\theta=\hat{\theta}} \xrightarrow[n\to\infty]{c.s.} \mathbb{E}_{Y\sim f}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\bigg|_{\theta=\theta_0}\right]$$

Si el modelo de la verosimilitud es correcto, entonces $f(Y) = p(Y|\theta^\star)$ para algún $\theta^\star \in \Theta$. Y, por lo tanto, la divergencia de Kullback-Leibler se puede escribir como

$$KL(\theta) = \mathbb{E}_{Y\sim f}\left[\log\left(\frac{p(Y|\theta^\star)}{p(Y|\theta)}\right)\right]$$

Recordando que $KL(\theta) \geq 0$, podemos verificar fácilmente que $KL(\theta^\star) = 0$ y, por lo tanto, $\theta_0 = \theta^\star$. Es decir, el parámetro que minimiza la divergencia de Kullback-Leibler es el verdadero parámetro. Entonces

$$\mathbb{E}_{Y\sim f}\left[\left.\frac{d^2}{d\theta^2}\log p(Y|\theta)\right|_{\theta=\theta_0}\right] = -J(\theta_0)$$

Sabemos que, a medida que $n \to \infty$, la distribución se concentra en vecindarios cada vez más pequeños de θ_0, y la distancia $|\hat{\theta} - \theta_0|$ se acerca a cero

Por lo tanto, al considerar los términos de la serie de Taylor, sólo necesitamos concentrarnos en el término cuadrático, de donde tenemos que

$$
\begin{aligned}
p(\theta|\mathbf{Y}) &\dot{\propto} \exp\left\{-\frac{1}{2}(\theta-\theta_0)^2(nJ(\theta_0))\right\} \\
&= \exp\left\{-\frac{(\theta-\theta_0)^2}{2(nJ(\theta_0))^{-1}}\right\}
\end{aligned}
$$

Es decir, para n suficientemente grande,

$$
\theta|\mathbf{Y} \,\dot{\sim}\, \mathsf{Normal}\left(\theta_0, (nJ(\theta_0))^{-1}\right).
$$

$\blacksquare$

Retomando la serie de Taylor, observamos lo siguiente

$$\log p(\theta|\mathbf{Y}) - \log p(\hat{\theta}|\mathbf{Y}) = \frac{1}{2}(\theta - \hat{\theta})^2 \ \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{Y})\bigg|_{\theta=\hat{\theta}}$$

$$\Rightarrow -2\log\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} = (\theta - \hat{\theta})^2 \left[-\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right]_{\theta=\hat{\theta}}.$$

Por lo tanto, si θ es de dimensión k, entonces:

$$-2\log\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \ \dot{\sim}\ \chi^2_k$$

Sea $q^{1-\alpha}_{\chi^2_k}$ el cuantil de probabilidad $1-\alpha$ de la distribución χ^2_k, i.e.

$$\mathbb{P}\left(\chi^2_k \leq q^{1-\alpha}_{\chi^2_k}\right) = 1-\alpha.$$

Entonces

$$\mathbb{P}\left[-2\log\frac{p(\theta|\mathbf{Y})}{p(\theta|\hat{\mathbf{Y}})} \leq q^{1-\alpha}_{\chi^2_k}\right] = 1-\alpha$$

$$\Rightarrow \mathbb{P}\left[\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \geq \exp\left\{-\frac{q^{1-\alpha}_{\chi^2_k}}{2}\right\}\right] = 1-\alpha.$$

Es decir, aquella región de Θ,

$$R(\Theta) = \left\{ \theta : \frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \geq \exp\left\{ -\frac{q^{1-\alpha}_{\chi^2_k}}{2} \right\} \right\}$$

corresponde a una región de aproximadamente $1 - \alpha$ probabilidad posterior.

En el caso de que θ sea de dimension k, entonces la serie de Taylor se escribiría como:

$$\log p(\theta|\mathbf{Y}) = \log p(\hat{\theta}|\mathbf{Y}) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + \cdots$$

y

$$\theta|\mathbf{Y} \;\dot\sim\; \mathsf{Normal}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

donde

$$I(\hat{\theta}) = -\left.\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right|_{\theta=\hat{\theta}}$$

Modelo Normal con previa no informativa

Sean $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathsf{Normal}(\mu, \sigma^2)$. Al ser μ y σ parámetros de localización y escala, respectivamente. Sabemos que una previa no informativa está dada por

$$p(\mu, \log\sigma) \propto \mathbb{1}_{\mathbb{R}}(\mu)\mathbb{1}_{\mathbb{R}}(\log\sigma)$$

y la verosimilitud de la muestra observada es

$$p(\mathbf{Y}|\mu, \log\sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

Entonces,

$$\begin{aligned}\log p(\mathbf{Y}|\mu, \log\sigma) &= \text{constante} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\\ &= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i-\mu)^2.\end{aligned}$$

Luego,

$$\log p(\mu, \log\sigma|\mathbf{Y}) = \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Sea $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$, se sigue que

$$
\begin{aligned}
&\log p(\mu, \log\sigma|\mathbf{Y}) \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^n (y_i-\mu)^2 \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^n (y_i-\bar{y}+\bar{y}-\mu)^2 \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]
\end{aligned}
$$

$$\frac{\partial}{\partial \mu} \log p(\mu, \log \sigma | \mathbf{Y}) = \exp\{-2 \log \sigma\} n(\bar{y} - \mu)$$

$$\frac{\partial}{\partial \log \sigma} \log p(\mu, \log \sigma | \mathbf{Y}) = -n + \exp\{-2 \log \sigma\} \left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]$$

Sea $(\hat{\mu}, \log \hat{\sigma})$ el punto en que se maximiza la densidad posterior. Podemos obtener este punto al evaluar en $(\hat{\mu}, \log \hat{\sigma})$ las expresiones anteriores e igualando a cero. De donde obtenemos que $\hat{\mu}) = \bar{y}$ y

$$\begin{aligned} &- n + \exp\{-2 \log \hat{\sigma}\} [(n-1)s^2] = 0 \\ \Rightarrow &\log \hat{\sigma} = \log \left(\sqrt{\frac{n-1}{n}} s \right). \end{aligned}$$

Calculamos ahora las segundas derivadas para obtener la información observada (de Fisher)

$$\frac{\partial^2}{\partial \mu \partial \log \sigma} \log p(\mu, \log \sigma | \mathbf{Y}) = -2n \exp\left\{-2 \log \sigma\right\} (\bar{y} - \mu)$$

$$\frac{\partial^2}{\partial \mu^2} \log p(\mu, \log \sigma | \mathbf{Y}) = -n \exp\left\{-2 \log \sigma\right\}$$

$$\frac{\partial^2}{\partial (\log \sigma)^2} \log p(\mu, \log \sigma | \mathbf{Y}) = -2 \exp\left\{-2 \log \sigma\right\} \left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]$$

Entonces,

$$I(\hat{\mu}, \log \hat{\sigma}) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & 2n \end{pmatrix}.$$

Por lo tanto,

$$\mu, \log \sigma | \mathbf{Y} \mathrel{\dot{\sim}} \mathsf{Normal}\left(\begin{pmatrix} \hat{\mu} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{1}{2n} \end{pmatrix} \right).$$

Ejercicio

Sabemos que si, en vez de haber considerado $\mu, \log \sigma$, hubiéramos considerado μ, σ^2, entonces la previa no informativa está dada por

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \mathbb{1}_{\mathbb{R}}(\mu) \mathbb{1}_{(0,\infty)}(\sigma^2).$$

Demuestre que en este caso

$$\mu, \sigma^2 | \mathbf{Y} \dot{\sim} \mathsf{Normal}\left(\begin{pmatrix} \hat{\mu} \\ \tilde{\sigma}^2 \end{pmatrix}, \begin{pmatrix} \frac{\tilde{\sigma}^2}{n} & 0 \\ 0 & \frac{2}{n+2}\tilde{\sigma}^4 \end{pmatrix} \right),$$

donde

$$\tilde{\sigma}^2 = \frac{n-1}{n+2} s^2 = \frac{n}{n+2} \hat{\sigma}^2.$$

Un modelo no regular

Sea $Y|a, b \sim \mathsf{Uniforme}(a, b)$,

$$p(Y|a, b) = \frac{1}{b-a} \mathbb{1}_{(a,b)}(y).$$

Considere la transformación dada por

$$U = \frac{Y-a}{b-a} \Rightarrow Y = (b-a)U + a$$

y

$$\frac{dY}{dU} = b - a,$$

luego

$$p(U|a, b) = \mathbb{1}_{(0,1)}(u).$$

Es decir $U \sim \mathsf{Uniforme}(0,1)$, como la distribución de U no depende de a, b ni Y, entonces U es una cantidad pivotal, a es parámetro de localización y $b - a$ es parámetro de escala. Así, podemos proponer la previa no informativa:

$$p(a, b-a) \propto \frac{1}{b-a} \mathbb{1}_{(0,\infty)}(b-a) \mathbb{1}_{\mathbb{R}}(a).$$

Considerando la reparametrización $\phi(a, b-a) = (a, b)$, obtenemos la previa no informativa para los parámetros a y b dada por

$$p(a, b) \propto \frac{1}{b-a} \mathbb{1}_{(a,\infty)}(b) \mathbb{1}_{\mathbb{R}}(a).$$

La verosimilitud puede ser escrita como

$$p(Y|a,b) = \frac{1}{b-a}\mathbb{1}_{(-\infty,y)}(a)\mathbb{1}\,(b)$$

y la verosimilitud de una muestra observada estaría dada por

$$p(\mathbf{Y}|a,b) = \frac{1}{(b-a)^n}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b).$$

Luego, la distribución posterior está dada por

$$p(a,b|\mathbf{Y}) \propto \frac{1}{(b-a)^{n+1}}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b).$$

Al integrar se puede calcular la constante de proporcionalidad y se demuestra que la densidad posterior es

$$p(a,b|\mathbf{Y}) = n(n-1)\frac{\left(y_{(n)}-y_{(1)}\right)^{n-1}}{(b-a)^{n+1}}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b)$$

Regresión Bayesiana

Regresión (Normal-Inversa Gama)

Detalles sangrientos

Tarea 2

- Dejar tarea 2.