Tema 6: Introducción a la Inferencia Bayesiana

Conchi Ausín
Departamento de Estadística
Universidad Carlos III de Madrid
concepcion.ausin@uc3m.es

CESGA, Noviembre 2012

Contenidos

El Teorema de Bayes

Recordamos que...

Dados dos sucesos A y B,

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A)\Pr(B \mid A)\\ &= \Pr(B)\Pr(A \mid B)\end{aligned}$$

Teorema de Bayes

$$\boxed{\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}}$$

Ejemplo: He visto a una persona con bolso. Calcular la probabilidad de que fuera mujer. Suponer que el 50 % de las personas son mujeres, el 60 % de las mujeres llevan bolso y el 20 % de los hombres también.

Probabilidades a priori y a posteriori

Dada una hipótesis, H, sobre una población, la inferencia Bayesiana la actualiza una vez que se han observado datos mediante,

$$\boxed{\Pr(H \mid Datos) = \frac{\Pr(Datos \mid H)\Pr(H)}{\Pr(Datos)}}$$

donde:

- $\Pr(H)$ es la probabilidad a priori de que la hipótesis H sea cierta.
- $\Pr(H \mid D)$ es la probabilidad a posteriori de que la hipótesis H sea cierta, la probabilidad de que H sea cierta una vez que se han observado los datos.
- $\Pr(Datos \mid H)$ es la verosimilitud, es decir, la probabilidad de que haber observado esos datos si la hipótesis H es cierta.
- $\Pr(Datos)$ es la verosimilitud marginal, la probabilidad de que haber observado esos datos independientemente de que H sea cierta o no.

Pasos en la inferencia Bayesiana

Supongamos que estamos interesados en estimar un parámetro, θ, a partir de unos datos $\mathbf{x} = \{x_1, \ldots, x_n\}$.

Con la filosofía Bayesiana θ no es un valor fijo, sino una variable aleatoria. Los pasos esenciales son:

1. Fijar una distribución a priori para θ, que denotamos por $\pi(\theta)$, que exprese nuestras creencias sobre θ antes de observar los datos.
2. Datos los datos, $\mathbf{x}$, escoger un modelo estadístico que describa su distribución dado θ, es la verosimilitud $f(\mathbf{x} \mid \theta)$.
3. Usando el Teorema de Bayes, actualizar las creencias sobre θ calculando su distribución a posteriori:

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\, \pi(\theta)}{f(\mathbf{x})}$$

Pasos en la inferencia Bayesiana

La verosimilitud marginal o distribución marginal de los datos,

$$f(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\pi(\theta)d\theta$$

es una constante de integración que asegura que la distribución a posteriori de θ integre uno, no depende de θ.

Por tanto, esta constante no proporciona ninguna información adicional sobre la distribución a posteriori y a menudo se escribe,

$$\pi(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)\,\pi(\theta)$$

Esta expresión es la distribución a posteriori sin normalizar, que es proporcional a la verosimilitud multiplicada por la distribucin a priori.

Un intervalo creíble al 95 % para θ es simplemente un intervalo (a, b) tal que la probabilidad a posteriori de que θ esté en el intervalo es del 95 %.

Inferencia Bayesiana en variables binarias

Supongamos que tenemos una moneda y queremos estimar la probabilidad de obtener cara,

$$\Pr(X = cara \mid \theta) = \theta, \qquad \Pr(X = cruz \mid \theta) = 1 - \theta.$$

Imaginemos que nuestras creencias a priori sobre θ se pueden describir como una distribución uniforme, $\mathcal{U}(0,1)$, luego la distribución a priori de θ es,

$$\pi(\theta) = 1, \quad 0 < \theta < 1.$$

Para actualizar la distribución de θ, realizamos el experimento de tirar la moneda 12 veces y obtenemos 9 caras y 3 cruces. Con estos datos, $\mathbf{x} = \{x_1, \ldots, x_{12}\}$, la verosimilitud es,

$$f(\mathbf{x} \mid \theta) = \binom{12}{9} \theta^9 (1 - \theta)^3$$

Inferencia Bayesiana en variables binarias

Luego, la distribución a posteriori de θ es proporcional a,

$$\pi(\theta \mid \mathbf{x}) \propto \theta^9 (1-\theta)^3 ,$$

que es el núcleo de una distribución beta,

$$\pi(\theta \mid \mathbf{x}) = \frac{\theta^9 (1-\theta)^3}{\int_0^1 \theta^9 (1-\theta)^3 \, d\theta} = \frac{1}{\mathcal{B}(10,4)} \theta^{10-1}(1-\theta)^{4-1}$$

Luego, $\theta \mid \mathbf{x} \sim \mathcal{B}(10,4)$.

Recordamos que la densidad de una beta $\mathcal{B}(\alpha,\beta)$ es,

$$\pi(\theta \mid \alpha,\beta) = \frac{1}{\mathcal{B}(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

donde

$$\mathcal{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Inferencia Bayesiana en variables binarias

Teniendo en cuenta que la media de una $\theta \sim \mathcal{B}(\alpha, \beta)$ es,

$$E[\theta \mid \alpha, \beta] = \frac{\alpha}{\alpha + \beta},$$

entonces, dados los datos, la media a posteriori de θ es,

$$E[\theta \mid \mathbf{x}] = \frac{10}{14} = \frac{5}{7} = \frac{1}{7} \times \frac{1}{2} + \frac{6}{7} \times \frac{9}{12}.$$

Luego,

$$E[\theta \mid \mathbf{x}] = \frac{1}{7} E[\theta] + \frac{6}{7} \hat{\theta}_{MV}$$

donde $E[\theta] = 1/2$ es la media a priori de la $\mathcal{U}(0, 1)$ y $\hat{\theta}_{MLE} = 9/12$ es la estimación máximo verosímil.

Luego, la media a posteriori es una media ponderada de la media de nuestras creencias a priori y la estimación MV.

Inferencia Bayesiana en variables binarias

Observar que la distribución uniforme, $\mathcal{U}(0,1)$, es un caso particular de la densidad beta cuando $\alpha = \beta = 1$, de modo que la distribución a priori y la posteriori son distribuciones beta, se dice que son conjugadas.

De modo más general, si asumimos a priori una distribución beta $\mathcal{B}(\alpha, \beta)$,

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

y la verosimilitud es,

$$f(\mathbf{x} \mid \theta) \propto \theta^{k}(1-\theta)^{n-k},$$

la distribución a posteriori es una beta $\mathcal{B}(\alpha + k, \beta + n - k)$,

$$\pi(\theta \mid \mathbf{x}) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}.$$

Luego,

$$E[\theta \mid \mathbf{x}] = \frac{\alpha + k}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} E[\theta] + \frac{n}{\alpha + \beta + n} \hat{\theta}_{MV}$$

Inferencia Bayesiana en variables binarias

Ejemplo: El archivo de datos `birthwt` en la librería `MASS` de `R` incluye, en la variable `bwt`, el peso al nacer de 189 niños en Massachusetts en 1986. Se considera que un bebé tiene bajo peso al nacer si pesa menos de 2500 gramos. Calcular:

1. La probabilidad a posteriori de que la proporción de niños con bajo peso sea mayor del 27.5 %
2. Un intervalo creíble al 95 % para dicha proporción.
3. Comparar los resultados con los obtenidos de forma frecuentista.

Usando:

- Una distribución a priori no informativa: $\mathcal{B}(1,1)$.
- Una distribución a priori informativa: $\mathcal{B}(1,5)$ ó $\mathcal{B}(5,1)$.
- Una distribución a priori informativa que refleje un estudio anterior que dió lugar a un intervalo de confianza al 95 % para la proporción de niños con bajo peso entre el 10 % y el 20 %.

Inferencia Bayesiana en variables binarias

Si tomamos α y β cada vez más pequeños: $\alpha, \beta \to 0$, la media a posteriori de θ converge a $\hat{\theta}_{MV}$. Con esta elección, la distribución a priori sería:

$$\pi(\theta) \propto \frac{1}{\theta(1-\theta)},$$

que no es una función de densidad ya que $\int \pi(\theta) d\theta = \infty$, se dice que es una distribución a priori impropia.

Esta elección a priori es válida ya que da lugar a una distribución a posteriori propia: $\theta \mid \mathbf{x} \sim \mathcal{B}(k, n-k)$.

Las distribuciones impropias permiten no imponer información subjetiva a priori. Sin embargo, es importante verificar que la distribución a posteriori sea propia.

Ejemplo: Usar una distribución a priori impropia en el ejemplo anterior.

Comparación con resultados frecuentistas

Los resultados con inferencia Bayesiana y frecuentista son similares cuando:

1. Se usan distribuciones a priori objetivas.
2. El tamaño muestral es muy grande y la influencia de la distribución a priori es muy pequeña en comparación con la influencia de la verosimilitud.

Inferencia Bayesiana en variables normales

Suponemos ahora que estamos interesados en hacer inferencia para una población normal, $X \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$, con densidad,

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Para facilitar los cálculos, es mejor trabajar con la precisión $\phi = 1/\sigma^2$ en lugar de usar la varianza,

$$f\left(x \mid \mu, \phi\right) = \frac{\phi^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\phi}{2}\left(x-\mu\right)^2\right)$$

Dada una muestra de datos, $\mathbf{x} = \{x_1, \ldots, x_n\}$, de una variable $N(\mu, 1/\phi)$ y una distribución a priori sobre (μ, ϕ), el objetivo es obtener su distribución a posteriori.

Inferencia Bayesiana en variables normales

Dados los datos, $\mathbf{x}$, la verosimilitud de (μ, ϕ) es,

$$f(\mathbf{x} \mid \mu, \phi) = (2\pi)^{-n/2} \phi^{n/2} \exp\left(-\frac{\phi}{2}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right]\right)$$

donde $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Demostración

$$\begin{aligned}
f(\mathbf{x} \mid \mu, \phi) &= \prod_{i=1}^{n} \frac{\phi^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\phi}{2}(x_i-\mu)^2\right) \\
&= (2\pi)^{-n/2} \phi^{n/2} \exp\left(-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= (2\pi)^{-n/2} \phi^{n/2} \exp\left(-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\right) \\
&= (2\pi)^{-n/2} \phi^{n/2} \exp\left(-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)
\end{aligned}$$

Inferencia Bayesiana en variables normales

A priori podemos usar la distribución normal-gamma que es conjugada y viene dada por:

$$\mu \mid \phi \sim \mathcal{N}\left(m, \frac{1}{\alpha\phi}\right)$$
$$\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$$

o equivalentemente,

$$\begin{aligned}\pi(\mu,\phi) &= \pi(\mu \mid \phi) \times \pi(\phi) \\ &\propto \phi^{1/2} \exp\left(-\frac{\alpha\phi}{2}(\mu - m)^2\right) \times \phi^{\frac{a}{2}-1} \exp\left(-\frac{b}{2}\phi\right)\end{aligned}$$

Inferencia Bayesiana en variables normales

Dados los datos, $\mathbf{x}$, la distribución a posteriori es también una normal-gamma con:

$$\mu \mid \phi, \mathbf{x} \sim \mathcal{N}\left(m^*, \frac{1}{\alpha^* \phi}\right)$$
$$\phi \mid \mathbf{x} \sim \mathcal{G}\left(\frac{a^*}{2}, \frac{b^*}{2}\right)$$

donde,

$$\begin{aligned} m^* &= \frac{\alpha m + n\bar{x}}{\alpha + n}, \\ \alpha^* &= \alpha + n, \\ a^* &= a + n, \\ b^* &= b + (n-1)\, s^2 + \frac{\alpha n}{\alpha + n}(m - \bar{x})^2. \end{aligned}$$

Observar que m^* es una media ponderada de la media a priori de μ y su estimador MV.

Inferencia Bayesiana en variables normales

Demostración

$$\pi(\mu,\phi) \propto \pi(\mu,\phi)\, f(\mathbf{x} \mid \mu,\phi)$$

$$\propto \phi^{\frac{a}{2}-1} \exp\left(-\frac{\phi\left[b+\alpha(\mu-m)^2\right]}{2}\right) \times \phi^{\frac{n}{2}} \exp\left(-\frac{\phi\left[(n-1)s^2+n(\mu-\bar{x})^2\right]}{2}\right)$$

$$\propto \phi^{\frac{a+n-1}{2}} \exp\left(-\frac{\phi\left[b+(n-1)s^2+(\alpha+n)\mu^2-2(\alpha m+n\bar{x})\mu+\alpha m^2+n\bar{x}^2\right]}{2}\right)$$

$$\propto \phi^{\frac{1}{2}} \exp\left(-\frac{\phi}{2}\left[(\alpha+n)\left(\mu-\frac{\alpha m+n\bar{x}}{\alpha+n}\right)^2\right]\right)$$

$$\times\, \phi^{\frac{a+n}{2}-1} \exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+\alpha m^2+n\bar{x}^2-\frac{(\alpha m+n\bar{x})^2}{\alpha+n}\right]\right)$$

$$\propto \phi^{\frac{1}{2}} \exp\left(-\frac{\phi(\alpha+n)}{2}\left(\mu-\frac{\alpha m+n\bar{x}}{\alpha+n}\right)^2\right)$$

$$\times\, \phi^{\frac{a+n}{2}-1} \exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+\frac{\alpha n}{\alpha+n}(m-\bar{x})^2\right]\right)$$

Inferencia Bayesiana en variables normales

En la práctica, el interés está en la distribución marginal a posteriori de μ que viene dada por:

$$\mu \mid \mathbf{x} \sim \mathcal{T}_{a^*}\left(m^*, \frac{b^*}{a^*\alpha^*}\right)$$

donde $\mathcal{T}_a(\delta, \lambda)$ representa una distribución t de Student no estandarizada tal que:

$$\frac{\mathcal{T} - \delta}{\sqrt{\lambda}} \sim \mathcal{T}_a$$

donde $\mathcal{T}_a$ es una distribución t de Student standard con a grados de libertad.

Dados los datos, $\mathbf{x}$, un intervalo creíble al $(1-\alpha)\,\%$ para μ es:

$$m^* \pm t_{\alpha/2, a^*}\sqrt{\frac{b^*}{a^*\alpha^*}}$$

Inferencia Bayesiana en variables normales

Demostración

Sabemos que si $Z \sim N(0,1)$ y $V \sim \chi^2_\nu$, entonces:

$$\frac{Z}{\sqrt{V/\nu}} \sim \mathcal{T}_\nu$$

Dados los datos, se tiene que:

$$\sqrt{\alpha^* \phi}\,(\mu - m^*) \sim \mathcal{N}\,(0,1)\,,$$

y

$$b^* \phi \sim \mathcal{G}\left(\frac{a^*}{2}, \frac{1}{2}\right) \Longleftrightarrow b^* \phi \sim \chi^2_a,$$

luego,

$$\frac{\sqrt{\alpha^* \phi}\,(\mu - m^*)}{\sqrt{\frac{b^* \phi}{a^*}}} \sim \mathcal{T}_{a^*} \Longleftrightarrow \frac{\mu - m^*}{\sqrt{\frac{b^*}{a^* \alpha^*}}} \sim \mathcal{T}_{a^*}$$

Inferencia Bayesiana en variables normales

Ejemplo: El archivo de datos `crabs` en la librería `MASS` de `R` incluye, en la variable `CL`, la longitud del caparazón en milímetos de 200 cangrejos de la especie Leptograpsus variegatus observados en Fremantle, Australia. Asumiendo normalidad de dicha longitud, calcular:

1. La probabilidad a posteriori de que la media de la longitud sea mayor de 30 milímetros.
2. Un intervalo creíble al 95 % para dicha media.
3. Comparar los resultados con los obtenidos de forma frecuentista.

Usando una distribución normal-gamma a priori para (μ, ϕ):

- No informativa con $a = 0{,}01$, $b = 0{,}01$, $m = 0$ y $\alpha = 0{,}01$.
- Informativa que refleje un estudio anterior que con 50 datos dió lugar a un intervalo de confianza al 95 % para la media de la longitud del caparazón entre 20 y 30 milímetros.

Inferencia Bayesiana en variables normales

Si consideramos una distribución a priori impropia:

$$\pi(\alpha, \beta) \propto \frac{1}{\phi},$$

obtenemos que a posteriori:

$$\mu \mid \phi, \mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{1}{n\phi}\right)$$

$$\phi \mid \mathbf{x} \sim \mathcal{G}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

Y la distribución marginal de μ a posteriori es:

$$\mu \mid \mathbf{x} \sim \mathcal{T}_{n-1}\left(\bar{x}, \frac{(n-1)s^2}{(n-1)n}\right) \equiv \mathcal{T}_{n-1}\left(\bar{x}, \frac{s^2}{n}\right)$$

Inferencia Bayesiana en variables normales

Así, un intervalo creíble al $(1-\alpha)\,\%$ para μ es:

$$\bar{x} \pm t_{\alpha/2,n-1}\sqrt{\frac{s^2}{n}}$$

que es igual al intervalo de confianza clásico para una media de una distribución normal.

Ejemplo: Usar una distribución a priori impropia en el ejemplo anterior.

Predicción Bayesiana

Habitualmente, más que en los parámetros, el interés radica en la predicción de nuevos valores de la variable. Dados los datos observados, $\mathbf{x} = \{x_1, \ldots, x_n\}$, la distribución predictiva de una nueva observación es:

$$f(x_{n+1} \mid \mathbf{x}) = \int f(x \mid \theta)\pi(\theta \mid \mathbf{x})d\theta$$

Para una muestra de una población normal, $X \mid \mu, \phi \sim \mathcal{N}(\mu, 1/\phi)$, con una distribución normal-gamma a priori, se tiene que:

$$X_{n+1} \mid \mathbf{x} \sim \mathcal{T}_{a^*}\left(m^*, \frac{(\alpha^* + 1)b^*}{\alpha^* a^*}\right)$$

Predicción Bayesiana

Demostración

Podemos escribir $X_{n+1} = \mu + \epsilon_{n+1}$, donde,

$$\mu \mid \phi, \mathbf{x} \sim \mathcal{N}\left(m^*, \frac{1}{\alpha^* \phi^*}\right)$$
$$\epsilon \mid \phi, \mathbf{x} \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$$

Luego,

$$X_{n+1} \mid \phi, \mathbf{x} \sim \mathcal{N}\left(m^*, \left(1 + \frac{1}{\alpha^*}\right)\frac{1}{\phi}\right).$$

y

$$\phi \mid \mathbf{x} \sim \mathcal{G}(\alpha, \beta).$$

Por tanto,

$$X_{n+1} \mid \mathbf{x} \sim \mathcal{T}_{a^*}\left(m^*, \frac{(\alpha^* + 1)b^*}{\alpha^* a^*}\right)$$

Predicción Bayesiana

En particular, si se asume una a priori impropia, $\pi(\mu, \phi) \propto 1/\phi$, la distribución predictiva es,

$$X_{n+1} \mid \mathbf{x} \sim \mathcal{T}_{n-1}\left(\bar{x}, \frac{(n+1)s^2}{n}\right)$$

que coincide con la distribución de la predicción de una nueva observación desde el punto de vista clásico.

Ejemplo: Obtener la distribución predictiva para la longitud del caparazón de un nuevo cangrejo usando las distribuciones a priori consideradas en los ejemplos anteriores.

Comentarios finales

- Hemos visto cómo realizar inferencia bayesiana para variables binarias y normales. En estos casos existen distribuciones a priori conjugadas que facilitan la obtención de la distribución a posteriori.

- Existen otros modelos para los que existen distribuciones conjugadas, por ejemplo:
  - Para una población exponencial, $X \mid \theta \sim \mathcal{E}(\theta)$, asumiendo una gamma a priori para la tasa, $\theta \sim \mathcal{G}(\alpha, \beta)$.
  - Para una población Poisson, $X \mid \theta \sim \mathcal{P}(\theta)$, asumiendo una gamma a priori para la media, $\theta \sim \mathcal{G}(\alpha, \beta)$.
  - Para una población uniforme, $X \mid \theta \sim \mathcal{U}(0, \theta)$, asumiendo una Pareto a priori $\theta \sim \mathcal{PA}(\alpha, \beta)$.

- El uso de distribuciones cojugadas facilita los cálculos, pero no necesariamente dan lugar a las mejores elecciones a priori. Existen muchas otras alternativas como las distribuciones a priori de Jeffreys', o las distribuciones de referencia.

Comentarios finales

- Sin embargo, en la mayoría de los problemas en la práctica dado un modelo y una dsitribución a priori sobre los parámetros, la distribución a posteriori no es fácil de obtener analílitamente.
- Para abordar este problema, se pueden utilizar diferentes alternativas que vamos a estudiar en los siguientes temas:
    - Aproximaciones asintóticas.
    - Integración Monte Carlo.
    - Simulación Monte Carlo:
        - Con métodos directos.
        - Mediante cadenas de Markov.