Tema

2

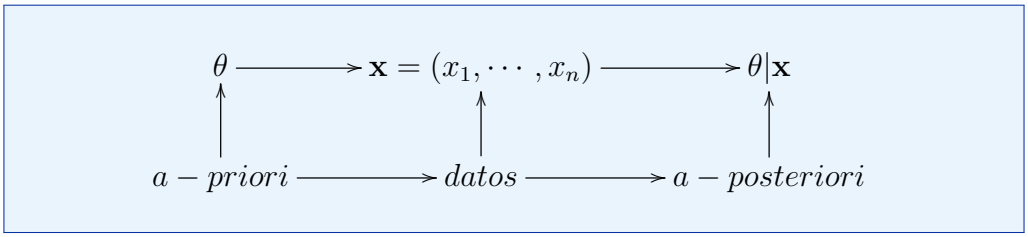
Conceptos Básicos en Inferencia Bayesiana

Índice

1. Estimación Bayesiana

1.1. Estimación Puntual

Desde el punto de vista Bayesiano, el cálculo de un estimador de un parámetro se construirá a partir de la distribución a posteriori. El gráfico muestra de manera sencilla cómo es el proceso.



Para justificar el uso de un estimador puntutal tenemos que definir algún concepto de optimalidad del estimador. Para ello usaremos la noción de función de pérdida.

Definición 1.1. *Una función $\mathcal{L} : \Theta \times \Theta \longrightarrow \mathbb{R}^+$ se dice que es una función de pérdida si cuantifica el error de estimación de $\widehat{\theta}$ con respecto a θ en algún sentido.*

Algunos tipos de funciones de pérdida son:

- Cuadrática: $\mathcal{L}(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$
- Norma L_1: $\mathcal{L}(\widehat{\theta}, \theta) = |\widehat{\theta} - \theta|$

El criterio de optimalidad que vamos a emplear será proponer como estimador aquel que minimice el error a posteriori esperado. A este estimador le llamaremos Estimador de Bayes.

Antes de demostrar la Proposición 1.2 necesitamos los siguientes lemas.

Lema 1.2. *Sea X una variable aleatoria no negativa con esperanza finita. Entonces se tiene que*

$$E(X) = \int_0^{+\infty} (1 - F_X(t))dt.$$

Demostración. Se tiene que por ser una integral impropia podemos expresarla de la siguiente manera:

$$E(X) = \lim_{m\to\infty} \int_0^m x dF(x).$$

Luego integrando por partes nos queda

$$\begin{aligned} E(X) &= \lim_{m\to\infty} \left(xF(x)|_0^m - \int_0^m F(x)dx \right) \\ &= \lim_{m\to\infty} \left(mF(m) - \int_0^m F(x)dx \right) \\ &= \lim_{m\to\infty} \left(m(F(m) - 1) + \int_0^m (1 - F(t))dt \right) \end{aligned}$$

Veamos ahora que $\lim_{m\to\infty} m(F(m)-1) = 0$, si la variable aleatoria tiene esperanza finita.

$$\begin{aligned} E(X) &= \lim_{m\to\infty} \left(\int_0^m x dF(x) + \int_m^\infty x dF(x) \right) \\ &\geq \lim_{m\to\infty} \left(\int_0^m x dF(x) + mP(X \geq m) \right) \\ &= E(X) + \lim_{m\to\infty} mP(X \geq m) \end{aligned}$$

Luego dicho límite es nulo ya que es el límite de una sucesión de números no negativos.

Por tanto, volviendo al desarrollo anterior, se obtiene que

$$E(X) = \int_0^\infty (1 - F(t))dt.$$

□

Ahora bien, supongamos que X es una variable aleatoria cualquiera. Definimos su parte positiva y negativa como

$$X^+ = \text{máx}\{0, X\} \text{ y } X^- = \text{máx}\{0, -X\}.$$

Es evidente que $|X| = X^+ + X^-$, luego $E(|X|) = E(X^+) + E(X^-)$. Es decir,

$$E(|X|) = \int_0^{+\infty} (1 - F_{X^+}(t))dt + \int_0^{+\infty} (1 - F_{X^-}(t))dt.$$

Mediante un cálculo sencillo se tiene que

$$F_{X^+}(t) = F_X(t) \text{ para } t > 0 \text{ y también } F_{X^-}(t) = 1 - F_X(-t) \text{ para } t > 0.$$

Por tanto,

$$E(|X|) = \int_0^{+\infty} (1 - F(t))dt - \int_{-\infty}^0 F(t)dt. \tag{1.1}$$

Veamos ahora la siguiente proposición.

Proposición 1.3. *El estimador de Bayes con respecto a la función de pérdida cuadrática es la media a posteriori, y con respecto a la función de pérdida de norma L_1 es la mediana.*

Demostración. Supongamos función de pérdida cuadrática. Entonces definimos la función:

$$g(\delta) = E\left((\theta - \delta)^2 | \mathbf{x}\right).$$

Donde δ es el estimador puntual de θ que obviamente depende de los datos $\mathbf{x}$ pero que no se denota por simplicidad. Luego,

$$g'(\delta) = 2(E(\theta|\mathbf{x}) - \delta).$$

Por tanto, $\delta = E(\theta|\mathbf{x})$.

Para el caso de la función de pérdida valor absoluto se tiene aplicando la ecuación (1.1) que

$$E(|\theta - \delta|) = \int_{\delta}^{\infty} [1 - P(\theta \leq t|\mathbf{x})]\, dt - \int_{-\infty}^{\delta} P(\theta \leq t|\mathbf{x}) dt.$$

Luego, derivando respecto a δ esta expresión e igualando a cero se obtiene que el valor óptimo δ^* es el que verifica

$$P(\theta \leq \delta^*|\mathbf{x}) = 1/2.$$

Es decir, $\delta^* = Med(\theta|\mathbf{x})$. □

1.2. Intervalos de credibilidad

Definición 1.4. *Dado un valor $\gamma \in (0,1)$, se define intervalo de credibilidad al $\gamma 100\,\%$ para $\theta \in \Theta \subset \mathbb{R}$ al intervalo (t_l, t_u) tal que*

$$\int_{t_l}^{t_u} f(\theta|\mathbf{x}) d\theta = \gamma,$$

donde $\mathbf{x} = (x_1, \cdots, x_n)$.

Los métodos que emplearemos para calcular los intervalos de credibilidad son los siguientes.

1. Para distribuciones unimodales conlleva calcular un intervalo de probabilidad que contenga a la moda (el máximo de la distribución a posteriori) tal que cualquier punto de este intervalo tenga un valor en la función de densidad mayor que cualquier otro punto que no esté en el intervalo. Este método se denomina HPDI de las siglas en inglés.

2. Proponer un intervalo cuya probabilidad coincida con la pedida que debe contener a la mediana. Se calcula repartiendo la misma probabilidad en la colas y por eso se denomina método de colas igualmente ponderadas.

3. Si la media existe, se propone un intervalo tal que la media sea el punto central.

2. Elección de la a priori

Este es el paso más importante de la Inferencia Bayesiana puesto que los cálculos posteriores dependen de ella. Este es el motivo de la mayor crítica por parte de la estadística frecuentista a la bayesiana.

2.1. Distribuciones conjugadas

Definición 2.1. *Sea $L(\theta) = f(\mathbf{x}|\theta)$ la verosimilitud del parámetro. Una clase $\mathcal{C}$ de distribuciones se llama conjugada respecto a la verosimilitud si para una a priori $\pi(\theta) \in \mathcal{C}$ la distribución a posteriori $\pi(\theta|\mathbf{x}) \in \mathcal{C}$ para todo $\mathbf{x}$.*

Ejemplo 2.2. *Sea $x_1, \cdots, x_n$ una muestra de una distribución Bernoulli de parámetro θ. Entonces se tiene que*

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}. \end{aligned}$$

Por tanto, la familia de distribuciones Beta es una familia conjugada respecto a la verosimilitud de una muestra Bernoulli.

2.2. Distribuciones impropias

En determinadas circunstancias, debido a la poca información disponible, puede ocurrir que la varianza sea muy grande lo cual puede originar que la función de densidad a priori sobre la varianza no sea tal, es decir, su integral no valga 1 y por tanto no pueda ser considerada como función de densidad de probabilidad.

Ejemplo 2.3. *Sea σ el parámetro desviación típica de un modelo distribucional. La única información que disponemos es que valores altos de varianza son muy poco probables. Una forma de modelar esta información es con la a priori dada por:*

$$f(\sigma) = \frac{1}{\sigma} I(\sigma > 0).$$

Sin embargo esta a priori es impropia ya que no es función de densidad puesto que su integral no vale la unidad.

Ejemplo 2.4. *Supongamos que $X_n \rightsquigarrow U[-n, n]$ para todo $n \geq 1$. Es decir su función de densidad viene dada por*

$$f_n(x) = \frac{1}{2n} I(x)_{[-n,n]} \quad \forall n \geq 1.$$

Pero si $n \to \infty$ la función de densidad tiende a cero y estaremos en un caso de distribución impropia.

2.3. Distribución a priori de Jeffrey

En el caso unidimensional, la distribución a priori no informativa de Jeffrey se basa en la información de Fisher dada por

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial log f(X|\theta)}{\partial \theta} \right)^2 \right].$$

Luego la distribución a priori de Jeffrey se define como

$$\pi(\theta) \propto I^{1/2}(\theta),$$

puediéndose dar el caso de ser una distribución impropia. Desde un punto de vista cualitativo, parece intuitivo pensar que los valores de θ para los que la cantidad de información de Fisher sea más grande serán más plausibles en la distribución a priori.

3. Distribuciones predictivas

Consideremos una muestra $\mathbf{x} = (x_1, \cdots, x_n)$ procedente de $f(x|\theta)$. Se pretende predecir una observación independiente $y \sim f(\cdot|\theta)$. La distribución predictiva se puede

calcular de dos maneras. Hay que tener en cuenta el proceso en la predicción. Es decir, como ya se tiene el conjunto de datos $\mathbf{x}$, la distribución a priori de θ se ha actualizado en la a posteriori $\pi(\theta|\mathbf{x})$ que a partir de este instante juega el papel de la *nueva* a priori. Por tanto, una forma sería aplicando el teorema de Bayes.

$$g(\theta|y,\mathbf{x}) = \frac{g(\theta|\mathbf{x})f(y|\theta,\mathbf{x})}{f(y|\mathbf{x})}.$$

Luego bastará despejar para obtener

$$f(y|\mathbf{x}) = \frac{f(y|\theta,\mathbf{x})g(\theta|\mathbf{x})}{g(\theta|y,\mathbf{x})}.$$

La segunda forma de calcular una densidad predictiva es mediante la ponderación con la densidad a posteriori:

$$f(y|\mathbf{x}) = \int_{\Theta} f(y|\theta,\mathbf{x})\pi(\theta|\mathbf{x})d\theta.$$

En ambos casos hay que tener cuidado con las constantes de proporcionalidad de las densidades a priori y a posteriori ya que su eliminación de estos métodos puede provocar densidades predictivas erróneas. Dichas constantes de proporcionalidad juegan un papel importante en estas densidades predictivas.

4. Factor de Bayes

Supongamos un contraste paramétrico de hipótesis estadísticas

$$\left.\begin{array}{l} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{array}\right\}$$

a partir de una muestra $\mathbf{x}$. La cuestión consiste en cuantificar la evidencia de las hipótesis con ese conjunto de datos. Para ello se define el factor de Bayes como el cociente:

$$FB_{01} = \frac{P(\mathbf{x}|H_0)}{P(\mathbf{x}|H_1)}.$$

Cuanto mayor sea el valor del factor de Bayes significa que hay más evidencia hacia la hipótesis nula que hacia la alternativa. Sin embargo es necesario realizar unos cálculos en términos de las a priori y la a posteriori para obtener FB_{01}.

Aplicando el Teorema de Bayes ya que en este caso *la causa* serían las hipótesis y *los efectos* los datos, se tiene que

$$P(\mathbf{x}|H_0) = \frac{P(\mathbf{x})P(H_0|\mathbf{x})}{P(H_0)},$$

de forma similar sería para la hipótesis alternativa. Luego si, las probabilidades a priori de las hipótesis son

$$f_0 = P(\theta \in \Theta_0) \text{ y } f_1 = P(\theta \in \Theta_1).$$

Después de observar una muestra, las probabilidades a posteriori de ambas hipótesis resultan ser

$$\alpha_0 = P(\theta \in \Theta_0|\mathbf{x}) \text{ y } \alpha_1 = P(\theta \in \Theta_1|\mathbf{x}).$$

El Factor de Bayes (FB) se puede calcular como l

$$FB_{01} = \frac{\alpha_0/\alpha_1}{f_0/f_1}.$$

La cuestión consiste en detectar que probabilidad cambia más en relación a la información que aportan los datos. Por ejemplo, la H_0 se verá mucho más creíble que H_1 si le afectan más los datos, es decir

$$\frac{P(H_0|\mathbf{x})}{P(H_0)} >> \frac{P(H_1|\mathbf{x})}{P(H_1)}.$$

Equivalentemente,

$$\frac{\alpha_0 f_1}{\alpha_1 f_0} >> 1.$$

Luego, el Factor de Bayes (FB) se puede interpretar como la plausibilidad a posteriori entre la plausibilidad a priori:

$$FB_{01} = \frac{\alpha_0/\alpha_1}{f_0/f_1}.$$

La clasificación de la evidencia hacia una hipótesis u otra se suele realizar mediante el uso de la Tabla de Jeffrey.

Ver Práctica-R-2

Factor de Bayes FB_{01}			Interpretación
	$>$	100	Evidencia extrema para H_0
30	–	100	Evidencia muy fuerte para H_0
10	–	30	Evidencia fuerte para H_0
3	–	10	Evidencia moderada para H_0
1	–	3	Evidencia anecdótica para H_0
	1		Sin evidencia
1/3	–	1	Evidencia anecdótica para H_1
1/10	–	1/3	Evidencia moderada para H_1
1/30	–	1/10	Evidencia fuerte para H_1
1/100	–	1/30	Evidencia muy fuerte para H_1
	$<$	1/100	Evidencia extrema para H_1

Cuadro 1: Tabla de puntos de corte para FB de Jeffrey