1. Estimación Bayesiana

La idea fundamental de la estimación bayesiana es que considera que el parámetro desconocido θ es una **variable aleatoria**. Como tal, θ tiene, antes de observar ningún dato, una función de densidad (o función de probabilidad si es una v.a. discreta) que representa los conocimientos previos o información inicial sobre el parámetro. Esta función de densidad se conoce con el nombre de **distribución a priori**, $f(\theta)$.

Esta distribución a priori se modifica basándonos en la información proporcionada por la m.a.s. $X_1, \ldots, X_n$. Esta información viene representada por la **función de verosimilitud**, que en estimación bayesiana denotaremos por $f(X_1, \ldots, X_n|\theta)$ para expresar el modelo del que provienen los datos condicionado a un valor genérico del parámetro desconocido.

Con estos dos ingredientes y el Teorema de Bayes (versión para v.a., ver glosario) actualizamos la información que tenemos sobre el parámetro θ, obteniendo la **distribución a posteriori** $f(\theta|X_1, \ldots, X_n)$:

$$f(\theta|X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n|\theta)f(\theta)}{\displaystyle\int_\Theta f(X_1, \ldots, X_n|\theta)f(\theta)d\theta}$$

El denominador no depende de θ y actúa como una constante normalizadora para que la integral $\displaystyle\int f(\theta|X)d\theta$ sea 1. Por lo tanto:

$$\text{A posteriori} \propto \text{A priori} \times \text{Verosimilitud}$$

$f(\theta)$ representa la verosimilitud relativa antes de observar la muestra.

$f(\theta|X_1, \ldots, X_n)$ representa la verosimilitud relativa después de observar la muestra.

<u>**Definición:**</u> Si las distribuciones a priori y a posteriori pertenecen a la misma familia de distribuciones, esa familia recibe el nombre de **familia conjugada** respecto de la distribución de la muestra de datos (respecto de la función de verosimilitud).

1.1. Estimador Bayes

Para formular inferencias con respecto a un parámetro θ se usa la distribución a posteriori, que es la a priori modificada una vez observada la muestra. Pero, ¿cómo se usa $f(\theta|X_1, \ldots, X_n)$ para obtener un estimador puntual de θ?

1. Podemos tomar como estimador el máximo, la moda, de la distribución a posteriori, que es el valor más probable, o cualquier otra característica como la mediana o la moda.

2. Podemos definir un criterio de optimalidad y deducir el estimador a partir de él. Definimos una función de pérdida $g(\hat{\theta}, \theta)$ que evalúe la pérdida que obtenemos usando $\hat{\theta}$ cuando el verdadero valor es θ. Escogeríamos entonces el estimador que minimiza la esperanza a posteriori de esa función de pérdida.

Por ejemplo, si utilizamos la función de pérdida cuadrática o error cuadrático, $g(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, el estimador que minimiza esta pérdida con respecto a la distribución a posteriori de θ es la media de dicha distribución, si utilizamos la función de pérdida valor absoluto, $g(\hat{\theta},\theta) = |\hat{\theta}-\theta|$, entonces debemos usar como estimador de θ la mediana de la distribución a posteriori, si maximizamos la verosimilitud a posteriori, debemos usar la moda, etc.

Los estimadores Bayes son consistentes, se aproximan al verdadero valor del parámetro a medida que aumenta el tamaño muestral. Para n grande, se aproximan mucho a los EMV (estimadores obtenidos por el método de máxima verosimilitud).

En general, nosotros tomaremos como estimador puntual bayesiano del parámetro θ, la media o esperanza de la distribución a posteriori para θ, es decir:

$$\hat{\theta} = E[\theta|X_1,\ldots,X_n]$$

CUADRO DE FAMILIAS CONJUGADAS

$\theta \sim$ A priori	$X\|\theta \sim$ Población	$\theta\|X_1,\ldots,X_n \sim$ A posteriori
$\mathcal{B}eta(\alpha,\beta)$	$\mathcal{B}\mathrm{er}(\theta)$	$\mathcal{B}\mathrm{eta}(\alpha+\sum_{i=1}^{n} X_i,\ \beta+n-\sum_{i=1}^{n} X_i)$
	$\mathcal{B}\mathrm{in}(m,\theta)$	$\mathcal{B}\mathrm{eta}(\alpha+\sum_{i=1}^{n} X_i,\ \beta+mn-\sum_{i=1}^{n} X_i)$
$\mathcal{G}amma(a,p)$	$\mathcal{P}(\theta)$	$\mathcal{G}amma(a+n,\ p+\sum_{i=1}^{n} X_i)$
	$\mathcal{E}xp(\theta)$	$\mathcal{G}amma(a+\sum_{i=1}^{n} X_i,\ p+n)$
$\mathcal{N}(\mu,\sigma_0)$	$\mathcal{N}(\theta,\sigma)$ con σ conocida	$\mathcal{N}\left(\frac{\sigma^2\mu+n\sigma_0^2\bar{X}}{\sigma^2+n\sigma_0^2}, \sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}}\right)$

Ejemplo 1: Supongamos que la v.a. X sigue una distribución de Bernoulli, $X \sim \mathcal{B}\mathrm{er}(\theta)$ con el parámetro θ desconocido. A su vez, la distribución a priori sobre θ es $f(\theta) \sim$

$Beta(\alpha,\beta)$. Seleccionando una m.a.s. de X, $X_1,\ldots,X_n$, calcular la distribución a posteriori $f(\theta|X_1,\ldots,X_n)$. A partir de esa distribución a posteriori, determinar un estimador Bayes para el parámetro.

$$X \sim \mathcal{B}\text{er}(\theta), \quad P(X=x|\theta)=\theta^x(1-\theta)^{1-x}, \quad x\in\{0,1\}$$

La función de verosimilitud será:

$$P(X_1=x_1,\ldots,X_n=x_n|\theta)=\prod_{i=1}^{n}P(X=x_i|\theta)=\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

La distribución a priori sobre θ sabemos que es:

$$\theta \sim Beta(\alpha,\beta), \quad f(\theta)=\frac{1}{\mathcal{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{para} \quad 0\le\theta\le 1$$

Aplicando el Teorema de Bayes calculamos la distribución a posteriori:

$$\begin{aligned} f(\theta|x_1,\ldots,x_n) &= \frac{P(x_1,\ldots,x_n|\theta)f(\theta)}{\int_\theta P(x_1,\ldots,x_n|\theta)f(\theta)d\theta} \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{\mathcal{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{\mathcal{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &= \frac{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}{\int_0^1 \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}d\theta} \end{aligned}$$

Utilizando la definición de la función $\mathcal{B}$eta,

$$\mathcal{B}(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx, \quad \text{para } \alpha>0,\beta>0$$

Obtenemos la función de densidad de la distribución a posteriori, que es una $Beta(\sum x_i+\alpha, n-\sum x_i+\beta)$.

$$f(\theta|x_1,\ldots,x_n)=\frac{1}{\mathcal{B}(\sum x_i+\alpha,n-\sum x_i+\beta)}\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$$

Podemos utilizar como estimador Bayes, la media de la distribución a posteriori, que en este caso es:

$$\hat{\theta}=\frac{\sum x_i+\alpha}{\alpha+\beta+n}$$

Ejemplo 2: Comprobar que la familia gamma es conjugada respecto a la distribución de Poisson para el caso de $n=1$.

Dist. a priori	Verosimilitud	Dist. a posteriori
$\theta \sim \gamma(a,p)$	$X|\theta \sim \mathcal{P}(\theta)$	$\theta|x \sim$?

$$f(\theta|x) = \frac{f(x,\theta)}{f_X(x)} = \frac{P(x/\theta)f(\theta)}{\displaystyle\int_{-\infty}^{\infty} P(x/\theta)f(\theta)d\theta} = \frac{e^{-\theta}\dfrac{\theta^x}{x!}\dfrac{a^p}{\Gamma(p)}\theta^{p-1}e^{-a\theta}}{\displaystyle\int_0^{\infty}\left(e^{-\theta}\frac{\theta^x}{x!}\frac{a^p}{\Gamma(p)}\theta^{p-1}e^{-a\theta}\right)d\theta}$$

Desarrollamos el denominador:

$$\int_0^{\infty} e^{-\theta}\frac{\theta^x}{x!}\frac{a^p}{\Gamma(p)}\theta^{p-1}e^{-a\theta}d\theta = \frac{a^p}{\Gamma(p)x!}\int_0^{\infty} e^{-(1+a)\theta}\theta^{x+p-1}d\theta = \frac{a^p}{\Gamma(p)x!}\frac{\Gamma(x+p)}{(1+a)^{x+p}}$$

donde hemos utilizado la siguiente propiedad de la función gamma:

$$\int_0^{\infty} x^{p-1}e^{-ax}dx = \frac{\Gamma(p)}{a^p} \text{ siendo } a = b + ic \text{ con } b > 0$$

Así, sustituyendo obtenemos la función de densidad a posteriori:

$$f(\theta|x) = \frac{\dfrac{e^{-(1+a)\theta}\theta^{x+p-1}a^p}{x!\Gamma(p)}}{\dfrac{a^p}{x!\Gamma(p)}\dfrac{\Gamma(x+p)}{(1+a)^{x+p}}} = \frac{(1+a)^{x+p}}{\Gamma(x+p)}\theta^{x+p-1}e^{-(1+a)\theta}$$

que es la función de densidad de una Gamma de parámetros $1 + a$ y $x + p$, es decir, $\theta|x \sim \gamma(1 + a, x + p)$.

Observación 1: Si operamos algo más en el denominador, obtenemos:

$$f_X(x) = \frac{a^p}{\Gamma(p)x!}\frac{\Gamma(x+p)}{(1+a)^{x+p}} = \cdots = \binom{x+p-1}{x}\left(\frac{a}{1+a}\right)^p\left(\frac{1}{1+a}\right)^x \sim BN\left(p, \frac{a}{1+a}\right)$$

Así, la marginal de X (distribución predictiva de X a priori) es una distribución binomial negativa con un número fijo de éxitos igual a p y probabilidad de éxito $\dfrac{a}{1+a}$. Recordemos que en una distribución binomial negativa, la variable aleatoria X representa el número de fallos antes del n-ésimo éxito, en este caso, el p-ésimo (n= p).

Observación 2: Si consideramos una m.a.s. de tamaño n, es decir, $(X_1, \ldots, X_n)$, la función de verosimilitud es:

$$P(\vec{x}|\theta) = P(X_1 = x_1, \ldots, X_n = x_n/\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} = e^{-n\theta}\frac{\theta^{\sum x_i}}{\prod x_i!}$$

Sustituyendo en la expresión

$$f(\theta|\vec{x}) = \frac{P(\vec{x}|\theta)f(\theta)}{\displaystyle\int_{-\infty}^{\infty} P(\vec{x}|\theta)f(\theta)d\theta}$$

y operando igual que antes, obtenemos $\theta|\vec{x} \sim \gamma(a + n, p + \sum_{i=1}^{n} x_i)$.

Observación 3: Ls distribución predictiva de X a posteriori, es decir, la distribución de probabilidad para una nueva observación de X, x_n, después de haber observado la muestra $(X_1, \ldots, X_n)$, es:

$$
\begin{aligned}
f(x_n|x_1,\ldots,x_n) &= \int_0^\infty f(x_n|\theta,\mathbf{x})f(\theta|\mathbf{x})d\theta \\
&= \int_0^\infty f(x_n|\theta)f(\theta|\mathbf{x})d\theta \\
&= \int_0^\infty \left[\frac{\theta^{x_n}e^{-\theta}}{(x_n)!}\right]\left[\frac{(a+n)^{p+\sum x_i}}{\Gamma(p+\sum x_i)}\right] e^{-(a+n)\theta}\theta^{p+\sum x_i-1}d\theta \\
&= \left[\frac{1}{(x_n)!}\right]\left[\frac{(a+n)^{p+\sum x_i}}{\Gamma(p+\sum x_i)}\right]\int_0^\infty e^{-(a+n+1)\theta}\theta^{x_n+p+\sum x_i-1}d\theta \\
&= \left[\frac{1}{\Gamma(x_n+1)}\right]\left[\frac{(a+n)^{p+\sum x_i}}{\Gamma(p+\sum x_i)}\right]\left[\frac{\Gamma(x_n+p+\sum x_i)}{(a+n+1)^{x_n+p+\sum x_i}}\right] \\
&= \left[\frac{\Gamma(x_n+p+\sum x_i)}{\Gamma(x_n+1)\Gamma(p+\sum x_i)}\right]\left(\frac{a+n}{a+n+1}\right)^{p+\sum x_i}\left(\frac{1}{a+n+1}\right)^{x_n}
\end{aligned}
$$

Que es una $BN(p+\sum x_i, \frac{a+n}{a+n+1})$. La media de esta distribución predictiva a posteriori es la misma que la de la distribución a posteriori, pero su varianza es mayor.