Introducción a la Probabilidad Bayesiana. Parte 4, Estimación de parámetros y la distribución Gaussiana.

Martín Onetto

El viejo péndulo

Una amiga quiere conocer el período T del péndulo de un querido reloj viejo que tiene en su pared. Para estimarlo decide hacer 100 mediciones con un cronómetro en el punto de menor altura y mayor velocidad, y usarlos para hacer la inferencia. Ella dice que en general error tanto por demás que por de menos en la misma cantidad σ. Ella confía en que sus mediciones en promedio valen el valor real"del período T de su péndulo querido. Denotando como I a esta información y a los datos $D = \{t_i\}$ con $i = 1, ..., 100$, una forma de calcular el valor de T es el proponer una *likelihood*:

$$P(D|T, \sigma, I) = P(\{t_i\}|T, \sigma, I) = \prod_i P(t_i|T, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t_i - T)^2}{2\sigma^2}\right\} \quad (1)$$

donde σ es el parámetro que nos habla de la incerteza del método de medición y lo tomaremos como sabido e igual para todos los datos. El hecho que escribamos a la probabilidad de cada valor de t_i de manera independiente viene se corresponde con que el resultado de una medición no aporta información al resultado de otra por lo tanto la probabilidad de $P(t_i|t_j, T, \sigma, I) = P(t_i|T, \sigma, I)$ y, como vimos anteriormente, el resultado de la probabilidad de todas las mediciones a la vez resulta en el producto de cada una de ellas. La propuesta de *likelihood* para cada t_i se conoce como distribución Gaussiana y sólo usa la información de que existe un valor promedio de mi medición T, y una dispersión característica σ. Es más, es la distribución que menos información asume de un sistema considerando que solo depende de esas variables. (Para profundizar en esto recomiendo los capítulos de máxima entropía del libro de Jaynes o de Gregory que están en la bibliografía del curso.)

Luego, para representar un estado de desconocimiento sobre T total vamos a proponer una prior para T constante, con lo cual la distribución *posterior* de T resulta:

$$P(T|D, \sigma, I) \propto P(D|T, \sigma, I)P(T|I) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t_i - T)^2}{2\sigma^2}\right\} \times 1 \quad (2)$$

Es cuestionable tomar una distribución prior que le de cualquier valor posible a T, pero en una primera instancia lo justificaremos diciendo que las inferencias que hagamos con esta prior, que es constante en todos sus valores, serán equivalentes a otra que esté restringida entre valores "razonables", pero que a su vez también es chata en un entorno del valor promedio de los datos. Retomaremos esta discusión más adelante.

Para sacar conclusiones de la distribución *posterior* $P(T|D,\sigma,I)$ de la expresión 2 vamos a trabajar la productoria como:

$$\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t_i-T)^2}{2\sigma^2}\right\} = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{\sum_i \frac{-(t_i-T)^2}{2\sigma^2}\right\} \tag{3}$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (t_i-\bar{t})^2 + n(\bar{t}-T)^2\right\} \tag{4}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2/n}(\bar{t}-T)^2\right\} \tag{5}$$

donde tomamos como n al número total de mediciones, a $\bar{t} = \frac{\sum_i t_i}{n}$, y sólo nos interesan los factores que tienen dependencia funcional con T. Así entonces resulta la distribución posterior final con su correcta normalización:

$$P(T|D,\sigma,I) = \left(2\pi\frac{\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2/n}(\bar{t}-T)^2\right\} \tag{6}$$

Es interesante ver que **la distribución que obtuvimos para T es también Gaussiana, pero centrada en el promedio de los datos $\bar{y}$ y con una dispersión o ancho $\frac{\sigma}{\sqrt{n}}$.**

Es fácil confundir los roles de cada uno de los objetos en juego en el procedimiento que hicimos. Recordemos que, en un principio a partir de nuestra información y modelo construimos la distribución *likelihood* de los datos como el producto de gaussianas para cada t_i. Cada una de estas distribuciones gaussianas está caracterizada por los mismos parámetros, un valor central o parámetro de ubicación que llamaremos T y una dispersión respecto de este valor central σ. Por otro lado, cuando consideramos la distribución posterior $P(T|D,\sigma,I)$ debemos observarla como una función de los valores y resulta que también es una distribución gaussiana. Sin embargo, ahora en este caso los parámetros que la gobiernan provienen de los datos, el parámetro central o de ubicación resulta el promedio de los datos $\bar{y}$ y la dispersión o ancho característico es proporcional a σ en un factor $\frac{1}{\sqrt{n}}$.

Valor estimado

Si mi amiga luego de hacer todo el procesamiento de sus mediciones quisiera simplificar el resultado de la distribución en un número o en un rango. ¿Cómo lo hace? Los usos y costumbres de la literatura proceden en calcular:

$$E[X] = \int x f(x) dx \tag{7}$$

$$V[X] = E[X^2] - E[X]^2 \tag{8}$$

conocidas como *esperanza* o valor medio para E, y varianza V. Para el caso de la distribución posterior de nuestra inferencia tenemos $E[T|D, \sigma, I] = \bar{t}$ y $V = \frac{\sigma}{n}$. El valor estimado de T es:

$$(T)_{est} = E[T] \pm \sqrt{V[T]} = \bar{t} \pm \frac{\sigma}{\sqrt{n}} \tag{9}$$

esta estimación para el caso de la gaussiana nos da un intervalo del 66 % del área que abarca la distribución. Sin embargo, hay infinitos intervalos que tienen contienen este área, ¿qué tiene de especial?. Para el caso de la gaussiana, este es el intervalo más pequeño que contiene ese área. Para cualquier otro valor de área que quisiéramos contemplar, el intervalo más chico posible será aquel que sea simétrico y centrado en el valor medio. En nuestro ejemplo con $n = 100$ el ancho de la distribución de T se volvió $\sqrt{n} = 10$ veces más chico, lo cual nos habla de qué tan definida queda nuestra confianza sobre unos pocos valores de T a medida que incluímos datos. Esto podemos interpretarlo como que cada vez que incluímos datos estamos sacando grados de libertad a los posibles valores que puede tomar T tal que exista probabilidad de que hayan sucedido esos datos, dejándonos sólo unos pocos valores de T como candidatos.

Una prior informativa

Consideremos ahora que la prior en lugar de ser una distribución plana, está centrada en un valor T_0 y con una dispersión de $\sigma/3$, es decir 3 veces más definida que en las mediciones. Si usamos una gaussiana para usar esta información previa, el cálculo de la posterior resulta:

$$\begin{aligned} P(T|D, \sigma, I) &\propto P(D|T, \sigma, I) P(T|I) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2/n}(\bar{t} - T)^2\right\} \times \left(2\pi\sigma^2\right)^{-3/2} \exp\left\{-\frac{1}{2\sigma^2/3}(T - T_0)^2\right\} \end{aligned} \tag{10}$$

En nuestro ejemplo, podemos interpretar esta prior como unas simples mediciones que se hicieron previamente que consideran el largo del péndulo y la gravedad para dibujar un número de T_0, éstas fueron comentadas a mi amida antes de procesar sus datos y las incorporó a la hora de hacer una inferencia completa. El hecho de formular de información de una prior con la misma forma funcional que la likelihood para el parámetro de interés es algo que ayuda al cálculo sin pérdida de información, y se la suele llamar *prior conjugada*. En nuestro ejemplo podemos además cuantificar cuanto información nos provee esta prior en relación a los datos de la siguiente manera, si hacemos el producto de ambas gaussianas considerando obtenemos:

$$\begin{aligned} P(T|D,\sigma I) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2/n}(\bar{t}-T)^2\right\} \times \left(2\pi\sigma^2\right)^{-3/2} \exp\left\{-\frac{1}{2\sigma^2/3}(T-T_0)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-(n+3)/2} \exp\left\{-\frac{1}{2\sigma^2/(n+3)}\left(T-\frac{\sum_i t_i + 3T_0}{n+3}\right)^2\right\} \end{aligned} \tag{11}$$

¿Qué tenemos acá? Si miramos sólo la distribución podemos descomponerla en la productoria de los n datos y tres otros factores de la forma:

$$\begin{aligned} &\left(2\pi\sigma^2\right)^{-(n+3)/2} \exp\left\{-\frac{1}{2\sigma^2/(n+3)}\left(T-\frac{\sum_i t_i + 3T_0}{n+3}\right)^2\right\} = \\ &\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(t_i-T)^2}{2\sigma^2}\right\} \times \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(T_0-T)^2}{2\sigma^2}\right\}\right]^3 \end{aligned} \tag{12}$$

Esto es idéntico a si tuvieramos tres mediciones extra idénticas en $T0$ y una distribución prior plana como la que usamos al comienzo!

En la estimación de T tenemos:

$$(T)_{est} = \frac{\sum_i t_i + 3T_0}{n+3} \pm \frac{\sigma}{\sqrt{n+3}} = \left(\bar{y} + \frac{3}{n+3}T_0\right) \pm \frac{\sigma}{\sqrt{n+3}} \tag{13}$$

Acá vemos que la información que nos da la prior de T en este caso corresponde a la de 3 mediciones de un mismo valor, la cual en última instancia no nos corre el valor de nuestro $(T)_{est}$ para el caso de que $n = 100$ como nuestro ejemplo a menos que T_0 ordenes de magnitud más grande que nuestro promedio $\bar{y}$. Lo cual diría que las distribuciones están centradas en valores muy diferentes y generaría sospechas sobre nuestros datos, nuestra distribución prior o nuestro modelo.

¿Cómo reportamos el valor de un parámetro en general? La aproximación Gaussiana.

En muchas situaciones no vamos a encontrarnos con la satisfacción de tener una distribución posterior conocida a la cuál le podemos calcular el valor medio y la varianza de forma simple, pero sabemos que sí va a suceder que las estimaciones de los parámetros se volverán angostas a medida de que tengamos más datos. En general si la distribución es unimodal, es decir que tiene un sólo máximo local,y en caso de tener un gran volumen de datos, una buena aproximación para la posterior será su desarrollo a segundo orden sobre el logaritmo evaluado en el máximo:

$$\log\left[P(\theta|D,I)\right] \sim \log\left[P(\theta=\theta_{max}|D,I)\right] + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\log\left[P(\theta|D,I)\right]|_{\theta=\theta_{max}}(\theta-\theta_{max})^2 \quad (14)$$

Esto se conoce como la aproximación de Laplace o gaussiana, dado que si exponenciamos el término a la derecha de la igualdad obtenemos una constante multiplicando a $\exp\left[\alpha(\theta-\theta_{max})^2\right]$. Sólo queda verificar que $\alpha = \frac{\partial^2}{\partial\theta^2}\log\left[P(\theta|D,I)\right]|_{\theta=\theta_{max}}$ es negativo, lo cual es inmediato sabiendo que estámos calculando la segunda derivada sobre un máximo. En caso de hacer esta aproximación podemos hacer la siguiente estimación del parámetro:

$$(\theta)_{est} = \theta_{max} \pm \frac{1}{\alpha} \quad (15)$$

en donde tenemos estimado el valor medio de la distribución por su máximo, y la $\sqrt{V}$ o ancho σ por la variable $\frac{-1}{\alpha}$, lo cual es exacto en el caso de que la distribución sea perfectamente gaussiana.