Introducción a la Probabilidad Bayesiana. Parte 3, Distribución Prior, Likelihood y Posterior

Martín Onetto

Resultados previos

$$P(A+B) = P(A) + P(B) - P(AB) \tag{1}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

$$\sum_i P(A_i) = 1 \tag{3}$$

$$P(B) = \sum_i P(A_i B) = \sum_i P(B|A_i)P(A_i) \tag{4}$$

Proposiciones de variables continuas

En general al hacer inferencia vamos a trabajar con variables que toman valores sobre un espacio continuo, ya sea determinando una constante universal o la tasa de éxito en un tratamiento. En esta situación, sea X la proposición de que la variable tome un valor $x \in \mathbb{R}$, ie $X = x$. La probabilidad de esta sentencia será:

$$P(X = x) = f(x)dx \tag{5}$$

donde $f(x)$ es ahora una distribución de probabilidad, conocida como *pdf*, probability density function. Sabemos que la probabilidad por definición debe estar normalizada por lo tanto tenemos que:

$$\int_{\mathbb{R}} f(x)dx = 1 \tag{6}$$

La marginalización en el caso continuo se traduce como:

$$P(X) = \int P(X, Y = y)dy = \int P(X|Y = y)P(Y = y)dy \tag{7}$$

Inferencia de Parámetros La notación que vamos a utilizar a la hora de hacer inferencia sobre parámetros de un modelo a partir de datos:

- D: Es la proposición que afirma que los datos tomaron ciertos valores.

- θ: Es la proposición de que afirma que los parámetros del modelo toman cierto valor
- I: EL modelos y toda la información que tenemos del experimento, cómo fue realizado y todo lo que sabemos de antemano sin utilizar D sobre los parámetros θ.

Con estas definiciones hacemos inferencia sobre los θ de la siguiente manera:

$$P(\theta|DI) = \frac{P(D|\theta, I)P(\theta|I)}{P(D|I)} \tag{8}$$

Cada uno de estos factores tiene nombre conocemos a $P(\theta|DI)$ como la distribución *posterior* de los parámetros, a $P(D|\theta, I)$ como la *likelihood* de los parámetros o también distribución de sampleo de los datos. A $P(\theta|I)$ la conocemos como la distribución *prior* de θ y finalmente a $P(D|I)$ se la entiende como una normalización de la *posterior* o también como *global likelihood*.

Los roles de estas distribuciones en la inferencia es el siguiente: la *likelihood* corresponde a la probabilidad de haber medido los datos D dado que el modelo es cierto y que tenemos un valor definido de θ. Luego, la distribución *prior* de θ representa el grado de certeza que tenemos para cada valor de θ del modelo, sin tener en cuenta los datos D. Por ejemplo si estamos haciendo inferencia sobre la masa del electrón, antes de hacer el experimento sabemos que no puede valer 1 gramo y a su vez, sí sabemos que es una cantidad estrictamente positiva. Por lo tanto una prior posible para θ es:

$$P(\theta|I) = \begin{cases} cte & \theta \in (0, 10\mathrm{e}^9 eV) \\ 0 & \text{en otro caso} \end{cases} \tag{9}$$

Esto es porque no yo no tengo mucha idea sobre cuanto vale la masa del electrón, pero alguien que viene haciendo este experimento hace décadas y conoce algunas cifras significativas, posee una prior mucho más angosta. Esto ya nos habla sobre una cualidad muy interesante sobre la inferencia y al probabilidad y es que son **subjetivas** al estado de información que tiene quien está haciendo el análisis.

OJO, según el Desiderata dos personas con la **misma** información no pueden llegar a conclusiones diferentes. Por lo que no tenemos ningún tipo de contradicción o problema, simplemente lo que vemos es que la extensión de la lógica se manifiesta en función de qué información tiene el que implementa estas herramientas.

La distribución *likelihood* $P(D|\theta I)$ es mide la probabilidad de haber medido esos datos dado nuestro modelo. En otras palabras la distribución está incluída en el modelo, la parte que suele denominarse *modelo estádistico*. Hay un zoológico de distribuciones posibles para modelar parámetros en función de qué sea lo que éstos representan, pero cómo ya mencionamos la información de los datos y nuestras inferencias no pueden llevarnos a conclusiones diferentes si consideramos distintas distribuciones bajo la misma información.

El producto de la prior con la likelihood constituye conceptualmente el proceso de actualización de información sobre el parámetro θ. Esta nueva distribución se manifiesta en la distribución *posterior*. En general este proceso es iterativo, podemos usar la posterior de hoy para usarla mañana como *prior* qué, junto a nuevas mediciones, volvemos a actualizar nuestra certeza sobre los valores que toma el parámetro.

El factor $P(D|I)$ que llamamos *global likelihood* consiste en una normalización sobre la distribución posterior. Eso lo podemos ver simplemente por el hecho de que funcionalmente es independiente de θ, por lo tanto corresponde al valor que tome la integral del numerador en la ecuación 8. Conceptualmente nos habla de la probabilidad de los datos dado el modelo teniendo en cuenta cualquier valor de los parámetros θ. Si $P(D|I)$ escribimos como:

$$P(D|I) = \int P(D\theta|I)d\theta = \int P(D|\theta I)P(\theta|I)d\theta \tag{10}$$

vemos que $P(D|I)$ es la distribución marginal de $P(D\theta|I)$, y por ende representa la probabilidad de haber medido los datos para cada valor del parámetro θ, pesado por la probabilidad *prior* que tenemos de ese valor de θ. En general, por el hecho de que $P(D|I)$ sea constante en θ hace que no cobre suma importancia a la hora trabajar con la posterior. Sin embargo, es la que va a cumplir un rol protagónico en la comparación de modelos que veremos más adelante. Sigamos con un ejemplo sobre como hacer uso de las distribuciones que definimos.

Un ejemplo, resultados de una moneda

Definamos las siguientes proposiciones como:

- I: "Mi amiga tiene una moneda que puede caer en cualquiera de sus dos lados, ambos son diferentes. No conozco ni cómo tira la moneda, ni si la moneda es justa, es decir si tiene tiene alguna cualidad que haga que caiga de un lado más que del otro. No hay ninguna influencia entre el resultado de una tirada de la moneda con la anterior ni con la siguiente".

- D: los datos corresponden al resultado de tirar 3 veces la moneda. Los fueron $\{c_1, c_2, s_3\}$, con c cara y s seca.

- θ: es el parámetro de nuestro modelo que nos dice cuanta confianza tenemos de que salga cara, $P(c|\theta I) = \theta$.

Si nos interesa saber que tan confiable es la moneda respecto a cada lado podemos calcular la distribución posterior de θ como:

$$P(\theta|DI) = \frac{P(D|I\theta)P(\theta|I)}{P(D|I)} \propto P(D|I\theta)P(\theta|I) \tag{11}$$

Identifiquemos cada término en nuestro ejemplo. En el $P(\theta|I)$ si usamos la información I vemos que no tenemos razón para asignar mayor probabilidad a ningún valor de θ por lo tanto la tomaremos plana, es decir

$$P(\theta|I) = \begin{cases} 1 & \theta \in (0,1) \\ 0 & \text{otro caso} \end{cases} \tag{12}$$

Luego, la *likelihood* la podemos expresar con lo siguiente

$$P(D|\theta I) = P(x_1 = c, x_2 = c, x_3 = s|\theta I) \tag{13}$$

donde x_i corresponde a la tirada i. En principio no sabemos cómo calcular esta probabilidad, pero con la regla del producto podemos decir que:

$$\begin{aligned} P(\{ccs\}|\theta I) &= P(x_1 = c, x_2 = c|x_3 = s, \theta I)P(x_3 = s|\theta I) \\ &= P(x_1 = c|x_2 = c, x_3 = s, \theta I)P(x_2 = c|x_3 = s, \theta I)P(x_3 = s|\theta I) \end{aligned} \tag{14}$$

Ahora bien, de la información I vemos que los resultados de cada tirada no tienen influencia sobre el resto, lo cual en el lenguaje de proposiciones corresponde a $x_j|x_i = x_j$ y por ende $P(x_j|x_iI) = P(x_j|I)$. Es así que la *likelihood* resulta en:

$$P(\{ccs\}|\theta I) = P(x_1 = c|\theta I)P(x_2 = c|\theta I)P(x_3 = s|\theta I) = \theta^2(1-\theta) \tag{15}$$

el término $P(x_3 = s|\theta I) = (1-\theta)$ es inmediato de la regla de la suma, y que $x_j \in c, s$.

La posterior de θ resulta entonces:

$$P(\theta|DI) \propto \theta^2(1-\theta) \tag{16}$$

La normalización en general para distribuciones de la forma $\theta^r(1-\theta)^{n-r}$ es un factor $\frac{(n+1)!}{r!(n-r)!}$, es así que la *posterior* escrita de manera completa queda:

$$P(\theta|DI) = \frac{3!}{2!}\theta^2(1-\theta) \tag{17}$$

¿Qué sucede si los datos fueran introducidos de a uno?

Imaginemos la situación que mi amiga mientras tira la moneda, me dice el resultado y yo voy actualizando sobre cada uno de ellos mi confianza en que salga cara, ie θ. Empezamos con $D_1 : \{x_1 = c\}$:

$$P(\theta|D_1, I) \propto P(D_1|\theta I)P(\theta|I) \propto \theta \times 1 \tag{18}$$

Hasta aquí entonces nuestra información sobre θ tomando sólo un dato es $P(\theta|D_1I) = 2\theta$. Si pensamos esto como un proceso de actualización de información, pasamos de una distribución plana para todos los valores de θ a una distribución lineal, donde mi confianza en cuánto vale θ crece linealmente con su valor. Definiendo $\tilde{I} = D_1I$ como mi nuevo estado de información la *prior* $P(\theta|\tilde{I})$ para la nueva tirada es $P(\theta|\tilde{I}) = 2\theta$. Haciendo uso de nuestra nueva información, calculamos nuevamente la *posterior* para la nueva tirada $D_2 = \{x_2 = c\}$

$$P(\theta|D_2, \tilde{I}) \propto P(D_2|\theta\tilde{I})P(\theta|\tilde{I}) \propto \theta \times 2\theta \tag{19}$$

Ahora volvimos a actualizar nuestra información sobre θ, $\hat{I} = D_2\tilde{I}$ donde su probabilidad vale $P(\theta|\hat{I}) = 3\theta^2$. Aquí vemos que el incremento en nuestra confianza de los valores que puede tomar θ pasó de ser lineal a cuadrático en valores positivos lo cual se condice

en que ya van dos veces que vemos que salió cara. En la siguiente tirada al salir seca, $D_3 = \{x_3 = s\}$, tenemos:

$$P(\theta|D_3, \hat{I}) \propto P(D_3|\theta\hat{I})P(\theta|\hat{I}) \propto (1-\theta) \times 3\theta^2 \tag{20}$$

El cual es el mismo resultado que obtuvimos considerando todos los datos juntos, ésto se debe a la simetría que nos impone la información sobre que las tiradas no tienen influencia lógica entre sí, pero influyen individualmente en nuestra información de θ.