Algoritmos de reconstrucción

Luis A. Escamilla

10 de julio de 2018

Resumen

Aquí va el abstract o resumen.

Estadística Bayesiana

Para entender cómo se hará la reconstrucción se necesita entender un poco de estadística bayesiana. En específico se necesita saber sobre cadenas de Markov y sobre el método de Cadenas de Markov Monte Carlo (MCMC por sus siglas en inglés). Por lo que se dará una breve introducción a la estadística bayesiana y sus conceptos antes de discutir MCMC.

Conceptos, definiciones y ejemplos

En estadística bayesiana las probabilidades se encuentran en la mente, no en el mundo. Esta frase nos da una pequeña noción sobre el cómo se trabaja con estadística bayesiana. Por ejemplo, supongamos que algún participante de un programa de televisión debe contestar una pregunta de opción múltiple, si hay 4 opciones, digamos que son incisos a, b, c y d, y la persona no tiene idea de cuál sea la respuesta correcta, entonces esa persona tiene 1/4 de probabilidades de escoger la opción correcta si elige al azar. Ahora supongamos que se le permite pedir ayuda a otra persona para responder la pregunta, si esta persona le dice “creo que es el inciso a” entonces el participante podría inclinarse en creer que el inciso a es la respuesta correcta, lo que haría que la probabilidad de a aumente en la mente del participante. Si dejan que pida ayuda a una tercera persona y ésta le responde también que cree que es el inciso a entonces el participante estará aún más seguro de que a es la respuesta correcta.

Analizemos este pequeño ejemplo que acabamos de plantear. El participante tenía 4 opciones iniciales pero no estaba seguro de cuál era la correcta, al recibir ayuda de la segunda persona recibió nueva información sobre el problema, lo que causó una inclinación hacia una de las 4 opciones, aumentando su probabilidad. Al hacerle la pregunta a la tercer persona y responder lo mismo que la segunda entonces aumentó aún más la probabilidad de que a fuese la respuesta correcta. Este pequeǹo ejemplo nos da la noción de un concepto importante para entender la estadística bayesiana: **al recibir nueva información sobre algún fenómeno para el que se tienen varias hipótesis que intenten explicar su comportamiento, cada una de ellas con su propia probabilidad de ser cierta, se deben actualizar las probabilidades de que alguna de estas hipótesis sea la correcta con base en la información que se recibió.** En el ejemplo de arriba las hipótesis son las 4 posibles respuestas, el fenómeno es la pregunta y la nueva información recibida es la respuesta de las 2 personas a las que se les pidió ayuda. Al pedirles ayuda se actualizaron las probabilidades de cada una de las 4 posibles respuestas pues la "nueva información" indicaba que a tenía mayor probabilidad de ser correcta que las otras 3.

Ahora veamos otro ejemplo en el que emplearemos números para poder dejar en claro la base de toda la estadística bayesiana: la **regla de Bayes.** Supongamos que tenemos una bolsa con 2 bolas dentro, una de ellas estamos completamente seguros de que es negra, pero la otra podría ser negra o blanca. Esta situación nos presenta dos casos a los que llamaremos **hipótesis** NN (dos bolas negras) y NB (una bola negra y una blanca). En principio ambos casos son igualmente probables, por lo que su probabilidad inicial es 0,5. Si lo acomodamos en una **caja de Bayes** quedará mejor representado:

Hipótesis	anterior	verosimilitud	h	posterior
NN	0.5			
NB	0.5			
Total:	1			

siendo $h = anterior \times verosimilitud$. Nuestras hipótesis tienen una cierta probabilidad inicial (anterior) que puede ser escogida dependiendo de las circunstancias. Si ahora procedemos a sacar una bola de la bolsa y resulta ser negra podremos empezar a calcular la *verosimilitud* (likelihood en inglés), que es un tipo muy especial de probabilidad. La verosimilitud se define como: probabilidad de haber calculado los datos suponiendo que la hipótesis fuese verdad. En este caso los "datos calculados" es el haber sacado una bola negra de la bolsa, si NN fuese la hipótesis correcta entonces la probabilidad de sacar una bola negra es 1, mientras que en NB es 0.5, por lo que la tabla queda:

Hipótesis	anterior	verosimilitud	h	posterior
NN	0.5	1	0.5	
NB	0.5	0.5	0.25	
Total:	1		0.75	

Pero aún nos falta calcular el *posterior*. El *posterior* es la "nueva probabilidad" dados los datos o información nueva y no es más que obtener la proporción de la *verosimilitud* de cada hipótesis con la suma de todas las verosimilitudes (básicamente lo normalizamos para obtener nuevas probabilidades que sumen 1). Entonces nuestra caja de Bayes está terminada:

Hipótesis	anterior	verosimilitud	h	posterior
NN	0.5	1	0.5	0.667
NB	0.5	0.5	0.25	0.333
Total:	1		0.75	1

Observemos que ahora NN tiene mayor probabilidad que NB pues hay mayor probabilidad de sacar una bola negra en NN en NB, haciendo que NN sea la hipótesis más probable luego de realizar la “medición”.

Este ejemplo nos ayuda a sentar las bases para definir la **regla de Bayes**:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{1}$$

siendo H la hipótesis, D los datos y:

- $P(H|D)$: probabilidad posterior. Qué tan probable es que H sea cierto dado que obtenemos D.
- $P(H)$: probablidad anterior. Qué tan probable es que H sea cierto antes de haber observado D.
- $P(D|H)$: verosimilitud. Qué tan probable es que observemos D si H es cierto.
- $P(D)$: verosimilitud marginal. Probabilidad de observar D sin saber si H es verdad o no.

y si generalizamos para N hipótesis:

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)}, \quad i = 1, 2, ..., N \tag{2}$$

$$P(D) = \sum_{i=1}^{N} P(H_i)P(D|H_i) \tag{3}$$

y esto podemos ponerlo en una caja de Bayes como:

Hipótesis	anterior	verosimilitud	h	posterior
H_1	$P(H_1)$	$P(D\|H_1)$	$P(H_1) \times P(D\|H_1)$	$P(H_1\|D)$
H_2	$P(H_2)$	$P(D\|H_2)$	$P(H_2) \times P(D\|H_2)$	$P(H_2\|D)$
...	...	...	...	...
Total:	1		$P(D)$	1

Estimación de parámetros

La estadística bayesiana también se puede utilizar para la estimación de parámetros de ciertas distribuciones. Para demostrarlo haremos un ejemplo.

Supongamos que un estudiante acaba de cambiarse de estado y está averiguando cómo ir a nueva universidad tomando el transporte público. Al investigar un poco se da cuenta de que tiene que tomar un camión en cierta calle y éste lo llevará a su universidad. En la primer semana toma 5 camiones, uno por día, y 2 de estos 5 camiones lo llevan a su universidad mientras que los otros 3 lo llevan más lejos y tiene que caminar más para poder llegar a la universidad. Este estudiante quiere entonces calcular el porcentaje de camiones que lo dejan cerca. Vamos a atacar este problema con estadística bayesiana.

Definamos a la proporción de camiones que lo dejan cerca como θ, dado que es una proporción entonces $0 \leq \theta \leq 1$. En principio θ puede tomar cualquier valor entre 0 y 1, pero por ahora lo restringiremos a tomar valores en intervalos de 0.1, esto nos da 11 hipótesis, cada una con probabilidad igual 1/11, escrbiendo esto en una caja de Bayes:

Hipótesis de θ	anterior $p(\theta)$	verosimilitud $p(x\|\theta)$	h $p(\theta)p(x\|\theta)$	posterior $p(\theta\|x)$
0	0.0909			
0.1	0.0909			
0.2	0.0909			
0.3	0.0909			
0.4	0.0909			
0.5	0.0909			
0.6	0.0909			
0.7	0.0909			
0.8	0.0909			
0.9	0.0909			
1	0.0909			
Total:	1			

Aquí se define x como el número de éxitos (número de camiones que si dejan al estudiante cerca de su escuela) de un total N de repeticiones. Si hay N repeticiones de un experimento "aleatorio" con probabilidad θ de que ocurra x veces el evento buscado entonces podemos expresar su función de probabilidad como una función binomial:

$$p(x|\theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}. \tag{4}$$

Esta función nos dice la distribución de probabilidad de x dado θ. Para este ejemplo $N = 5$ y $x = 2$. Esta distribución nos da entonces la verosimilitud para cada θ, teniendo la verosimilitud podemos calcular tanto h como la probabilidad posterior, entonces la caja de Bayes queda completa:

Hipótesis de θ	anterior $p(\theta)$	verosimilitud $p(x\|\theta)$	h $p(\theta)p(x\|\theta)$	posterior $p(\theta\|x)$
0	0.0909	0	0	0
0.1	0.0909	0.0729	0.0066	0.0437
0.2	0.0909	0.2048	0.0186	0.1229
0.3	0.0909	0.3087	0.0281	0.1852
0.4	0.0909	0.3456	0.0314	0.2074
0.5	0.0909	0.3125	0.0284	0.1875
0.6	0.0909	0.2304	0.0209	0.1383
0.7	0.0909	0.1323	0.0120	0.0794
0.8	0.0909	0.0512	0.0047	0.0307
0.9	0.0909	0.0081	0.0007	0.0049
1	0.0909	0	0	0
Total:	1		0.1515	1

Ahora analizemos este ejemplo ahora que tenemos la caja de Bayes completa. Como observamos tenemos una función de densidad de probabilidad (69), y queremos saber el valor del parámetro θ dados los datos x y N. En la tabla observamos que la probabilidad posterior mayor corresponde a (sin mucha sorpresa) $\theta = 0{,}4$. Entonces es seguro decir que, dada la información proporcionada, el valor del parámetro $\theta = 0{,}4$ es el que mejor se ajusta a la información. Esto tiene sentido, pues de los 5 camiones que pasaron, únicamente 2 lo dejaron en la escuela, y $\frac{2}{5} = 0{,}4$.

Podemos reescribir la regla de Bayes para la estimación de parámetros. Como el término $P(D)$ es un término de normalización podemos retirarlo para reescribir (69) como:

$$p(\theta|x) \propto p(\theta)p(x|\theta) \tag{5}$$

o

$$\textit{posterior} \propto \textit{anterior} \times \textit{verosimilitud} \tag{6}$$

Prueba de hipótesis (selección de modelo)

Como vimos en la sección anterior, para realizar estimación de parámetros se necesitan las probabilidades posteriores de los valores posibles del parámetro. El valor del parámetro con una probabilidad posterior mayor se considera que es el valor que mejor se ajusta con los datos u observaciones.

Para realizar una prueba de hipótesis (o selección de modelo) en estadística bayesiana sehace algo similar a una estimación de parámetros, pues los valores probables del parámetro son las hipótesis, pero un cierto intérvalo de valores también pueden ser una hipótesis. Un ejemplo rápido de este caso sería, utilizando el ejemplo anterior de los camiones, decir que tenemos 2 hipótesis posibles, la nula H_0 y la alternativa H_1. Las definimos como:

$$\begin{aligned} H_0 &: \quad \theta \leq 0{,}5 \\ H_1 &: \quad \theta > 0{,}5 \end{aligned}$$

por lo que tenemos dos posibilidades, o el valor de θ está de 0,5 hacia abajo o arriba de 0,5. Para obtener la probabilidad sólo tenemos que sumar las probabilidades posteriores de los valores independientes de θ según se ajusten a alguna de las dos hipótesis. Entonces

$$P(H_0|x) = P(\theta \leq 0{,}5|x) = 0{,}7467$$
$$P(H_1|x) = P(\theta > 0{,}5|x) = 0{,}2533.$$

En este ejemplo se cumple que $P(H_0|x) + P(H_1|x) = 1$ pues las hipótesis son mútuamente excluyentes (si ocurre H_0 es imposible que ocurra H_1).

La regla de Bayes se cumple para ambas hipótesis, entonces

$$P(H_0|x) = \frac{P(x|H_0)P(H_0)}{P(x)} \tag{7}$$

$$P(H_1|x) = \frac{P(x|H_1)P(H_1)}{P(x)} \tag{8}$$

y de aquí obtenemos la odds form de la regla de Bayes

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{P(H_0)}{P(H_1)}. \tag{9}$$

donde a $\frac{P(x|H_0)}{P(x|H_1)}$ se le llama **factor de Bayes**, el cuál nos dice qué tan probable es H_0 con respecto a H_1.

Cadenas de Markov

Para explicar la reconstrucción que se utilizará hace falta explicar un poco un proceso estocástico conocido como **Cadena de Markov**. Pero primero unas definiciones importantes:

- Proceso estocástico: proceso que pretende describir la evolución temporal de algún fenómeno aleatorio.

- Propiedad de Markov: propiedad de los procesos estocásticos que consiste en su ausencia de memoria, esto significa que la distribución de probabilidad de estados futuros del proceso depende únicamente del estado presente (ninguno de los estados anteriores al presente afectan esta distribución de eventos futuros, por esto se dice que tiene ausencia de memoria, pues no "recuerda"los pasos anteriores al actual o presente). Si un proceso tiene esta propiedad se le llama *proceso de Markov*.

Usando estas definiciones ahora podemos definir una cadena de Markov.

- Cadena de Markov: proceso de Markov que describe una secuencia de posibles eventos representándolos como "pasos" o "eslabones".

Ejemplos muy populares de procesos que pueden ser modelados por una cadena de Markov son los llamados *caminantes aleatorios*.

Dado que las cadenas de Markov son un proceso de Markov cada paso que da la cadena es indiferente a todos los anteriores excepto por el paso en el que se encuentre en ese momento.

Las cadenas de Markov pueden representarse como un diagrama de distintos estados. Pongamos un ejemplo, imaginemos que Pepito realiza alguna de las siguientes 3 actividades todos los días: correr, dormir o comer helado y por alguna razón un amigo suyo realizó una estadísica de su comportamiento. El amigo descubrió que si Pepito dormía, al día siguiente tenía una probabilidad de 0,6 de correr, una probabilidad de 0,2 de comer helado y una de 0,2 de volver a dormir. Si Pepito comía helado entonces era 0,7 de correr, 0,1 de comer helado de nuevo y 0,2 de dormir, y si corría entonces era 0,6 de volver a correr al día siguiente, 0,1 de dormir y 0,3 de comer helado. Esta información se puede representar de manera más fácil en un diagrama de un proceso de Markov como el de la Figura 1.

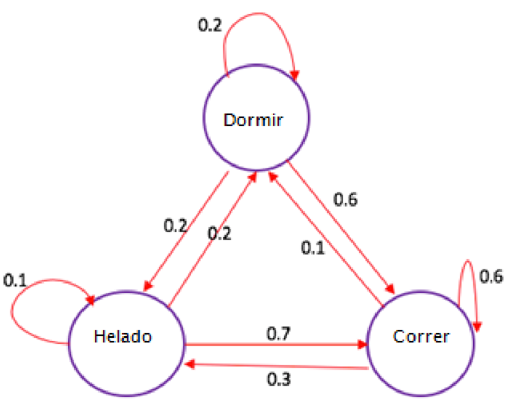

Figura 1: Diagrama que representa el proceso de Markov.

Si la cadena de Markov comenzase en el estado *dormir* podemos ver en el diagrama la probabilidad que tienen los 3 estados de ser el siguiente paso en la cadena.

Cadenas de Markov Monte Carlo (MCMC)

Se le llama *Monte Carlo* de manera general a alguna técniva computacional que utilice números aleatorios. Un tipo de Monte Carlo es MCMC, y un topo de método MCMC es el algoritmo Metropolis-Hastings. Este algoritmo nos da una manera de construir una cadena donde los valores de algún parámetro con probabilidades posteriores mayores aparezcan de manera más frecuente en la cadena que los valores con probabilidades posteriores menores, pero no hace que sea imposible que los valores con probabilidades menores ocurran, por lo que estarán presentes en la cadena pero con menos frecuencia.

De manera sencilla, el algoritmo funciona de la siguiente manera:

1. Se comienza en algún estado θ.

2. Se genera una propuesta θ' para el siguiente paso.

3. Si la $h' = posterior \times verosimilitud$ de θ' es mayor que la h del estado actual entonces se da el paso hacia θ' en la cadena y éste será el nuevo estado actual de la cadena, se reinicia el proceso con θ' como el nuevo estado inicial.

4. Si h' es menor entonces se propone una probabilidad mínima h'/h y se genera un número aleatorio entre 0 y 1, si el número generado es mayor o igual a h'/h entonces se da el paso y θ' será el nuevo estado actual en la cadena, se reinicia el proceso con θ' como nuevo estado inicial, pero si el número generado es menor que h'/h entonces se rechaza la propuesta θ' y se reinicia el proceso.

Si observamos bien el paso 4 del algoritmo podemos observar que h'/h es la ec. (74), por lo que el algoritmo Metropolis-Hastings es una forma de realizar comparaciones entre hipótesis. Cada hipótesis tiene una cierta probabilidad de ser cierta. Lo que se pretende lograr con MCMC es correr una cadena lo suficientemente larga como para...