Inferencia en el marco de la Estadística Bayesiana

Alfredo Mejía-Narváez

18 de agosto de 2017

Resumen

La estadística representa un conjunto de reglas diseñadas para expresar nuestro grado de conocimiento dados cierta evidencia y/o conocimiento previo, o al menos eso es para lo que ha sido concebida. Sin embargo, nuestra línea de razonamiento científico se ve forzada a desviarse un poco de este esquema, y la pregunta de interés científico real: ¿cuál es el grado de conocimiento sobre un evento, dados un conjunto de observaciones y conocimiento previo o prejuicios? se transforma en un *surrogate* en el que se supone que la hipótesis planteada es cierta: dado que tenemos conocimiento absoluto del estado del problema (e.g., el modelo con el cual describimos los datos con sus estadísticas correspondientes), ¿cuál es la verosimilitud de los datos?. La estadística Bayesiana presenta un esquema que permite responder naturalmente a la pregunta de interés científico, al mismo tiempo de la manera más objetiva posible. En este seminario les presentaré el esquema general de la inferencia bayesiana y algunas de sus aplicaciones.

1. Introducción: estadística básica

El concepto de probabilidad ha venido a responder una pregunta fundamental para el ciencia y otras áreas del conocimiento en general: ¿cuál es el grado de certidumbre que poseemos sobre un evento, dado conocimiento relevante de fondo?. En este sentido, una vez nuestro grado de certidumbre sobre un evento queda por sentado también así queda nuestro grado de intertidumbre. Por ejemplo, si sabemos que existe un 74 % de probabilidad de que llueva, sabemos que existe un 26 % de que no. Es en esta *completitud* en la que yace la mayor fortaleza del concepto de probabilidad y que es y debe ser particularmente apreciado en las ramas del conocimiento científico.

Las operaciones matemáticas sobre las cuales se construyen las bases de una teoría de probabilidad son bastante simples y están bien definidas. Hoy en día se conocen como las reglas de Cox y matemáticamente se resumen en dos ecuaciones:

$$1 = \Pr(\theta \,|\, I) + \Pr\left(\bar{\theta} \,\middle|\, I\right) \tag{1a}$$

$$\Pr(\theta_A, \theta_B \,|\, I) = \Pr(\theta_A \,|\, \theta_B, I) \times \Pr(\theta_B \,|\, I) \tag{1b}$$

En lenguaje cotidiano, la primera ecuación nos dice que dado nuestro grado de conocimiento (probabilidad) de conjunto exhaustivo y mutuamente excluyente de posibilidades, podemos asignar inmediatamente nuestro grado de desconocimiento (o ignorancia sobre el mismo conjunto). Ambas cantidades sumadas forman una *certeza*; la segunda ecuación nos dice que dado nuestro grado de conocimiento de θ_B y nuestro grado de conocimiento en θ_A condicionado en θ_B, podemos admitir que sabemos nuestro grado de conocimiento de θ_A y θ_B.

$\Pr(\theta \,|\, I)$: es nuestro grado de conocimiento sobre una hipótesis o premisa θ condicionado sobre I. En el sentido más general θ puede representar un vector, $\boldsymbol{\theta}$, en el espacio de hipótesis, $\boldsymbol{\Theta}$.

2. El teorema de Bayes

$\pi_N(\boldsymbol{\theta} \,|\, \{\mathcal{D}_i\}, I)$: es la distribución de probabilidad posterior y representa nuestro grado de conocimiento de la

El teorema de Bayes se desprende de estas dos simples relaciones y sostiene que

$$\pi_N\left(\theta \,|\, \{\mathcal{D}_i\}, I\right) = \frac{\mathcal{L}\left(\{\mathcal{D}_i\} \,|\, \theta, I\right) \times \pi_0\left(\theta \,|\, I\right)}{\Pr\left(\{\mathcal{D}_i\} \,|\, I\right)}, \tag{2}$$

donde

$$\Pr\left(\{\mathcal{D}_i\} \,|\, I\right) = \int_{\Theta} \pi_N\left(\theta \,|\, \{\mathcal{D}_i\}, I\right) d\theta.$$

$\Pr\left(\{\mathcal{D}_i\} \,|\, I\right)$: se conoce como la evidencia y es la distribución posterior marginalizada sobre el espacio de parámetros.

El poder del teorema de Bayes en sí yace en nuestra capacidad para asignar las distribuciones de probabilidad en el lado derecho de la igualdad, $\text{lik}\{\mathcal{D}_i\}\theta, I$ y $\pi_0\left(\theta \,|\, I\right)$, en la Ec. (2). Esto es usualmente cierto en el caso de la última distribución, pues esta refleja simplemente *nuestro grado de conocimiento (ignorancia y/o prejuicios) sobre problema previo a la obtención de los datos*, la primera distribución de probabilidades por otra parte, requiere un poco más de elaboración y conocimiento (probablemente también prejuicios) sobre los datos, pues representa la *plausibilidad de la hipótesis asumida a la luz de los datos*.

3. Inferencia bayesiana

La inferencia bayesiana es la técnica que consiste en calcular la distribución de probabilidad posterior, $\pi_N\left(\theta \,|\, \{\mathcal{D}_i\}, I\right)$ para hacer algún avance en nuestro grado de conocimiento sobre θ. El primer paso es construir la función que describe la probabilidad de haber hecho la observación suponiendo la hipótesis como correcta y la distribución prior que describe nuestro grado de conocimiento de la hipótesis antes de haber hecho la observación (véase el paréntesis A). Ambas distribuciones de probabilidad requieren de nuestro conocimiento del problema, pero en particular, la distribución *likelihood* solo se puede concebir correctamente si conocemos también los datos (e.g., los errores son Gaussianos, las observaciones son naturalmente poco probables, el número de posibilidades es bajo, etc.).

La elección de la distribución de probabilidad prior, por otra parte, representa un problema más complejo y que muchas veces es desestimado. Existen básicamente dos clases de distribuciones prior: las subjetivas que permiten que el científico introduzca sus prejuicios y las objetivas caracterizados generalmente por una distribución plana con la que solo se proporciona un rango de plausibilidad para la hipótesis en cuestión, permitiendo así que los datos influyan más sobre la distribución posterior. Por supuesto, como veremos a continuación, el efecto de una distribución prior subjetiva se disipará en la medida en que el volumen de datos sea más grande, i.e., el efecto de los datos dominara sobre nuestro grado de conocimiento de los mismos a través de la hipótesis planteada.

Es importante notar que no todo prior plano es objetivo, si por ejemplo, el rango de plausibilidad se restringe a uno más corto que el rango que incluye todas las posibilidades, entonces la distribución prior es subjetiva. De igual manera, una distribución prior no plana no necesariamente es subjetiva, pues una distribución prior pudo ser la distribución posterior de una inferencia anterior a la obtención de los datos actuales o incluso, pudo ser la posterior de un problema distinto. En estos casos decimos que la distribución prior está introduciendo dominio de conocimiento sobre el modelo.

4. Ejemplo: regresión lineal

Ahora vamos a hacer el ejercicio más vulgar y silvestre y, aún así uno de los más útiles si se hace bien: ajustar una línea recta a un conjunto de datos. Supongamos que dicho conjunto de datos está descrito por un arreglo independiente (i. e., dado) $\{x_i\}$ y un arreglo de medidas $\{y_i\}$

cuyas desviaciones estándar están dadas también $\{\sigma_i\}$. La intuición nos dice que por allí podría pasar una línea recta. La forma más objetiva de plantear el problema, desde el punto de vista probabilístico, consiste en construir un modelo generativo (**?**): una descripción estadística que permite reproducir los datos observados. Esto es esencialmente, construir la distribución de verosimilitud. Bajo la suposición de que los errores en $\{y_i\}$ son Gaussianos y no correlados, la verosimilitud es simplemente:

$$\mathcal{L}\left(\{y_i\}|\{x_i\},\{\sigma_i\},m,b\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - m\,x_i + b)^2}{2\sigma_i^2}\right].$$

En este punto el problema ya se ha convertido en un tecnicismo: ¿cómo encontrar los parámetros m y b, tal que la verosimilitud es máxima? La respuesta fácil (y muchas veces conveniente) a esa pregunta es simplemente minimizar la función de mérito χ^2. Uno podría, sin razón lógica aparente, simplemente ir directo a la inferencia Bayesiana y encontrar la distribución posterior $\pi_N\left(m,b\,|\{y_i\},\{x_i\},\{\sigma_i\}\right)$ usando distribuciones *prior* objetivas, por ejemplo. Sin embargo, ¿qué pasa si realmente no estamos seguros de que $\mathcal{L}\left(\{y_i\}|\{x_i\},\{\sigma_i\},m,b\right)$ realmente es un modelo generativo de todo el espacio de $\{y_i\}$? Y nos hacemos esta pregunta independientemente del la forma en que estén relacionados los parámetros de interés. Una forma de resolver este tipo de problemas es usando inferencia Bayesiana.

Bajo el esquema Bayesiano, las distribuciones *prior* no solo se usan para introducir conocimiento previo a la obtención de los datos, sino que al mismo tiempo se introducen incertidumbres (de acuerdo con la definición de probabilidad). Podemos usar esto a nuestro favor.

5. Efecto del prior

En esta sección veremos como la elección de la distribución prior puede afectar significativamente nuestra inferencia y como depende dicho efecto en la cantidad de los datos y en la forma del prior.

Para cuantificar el efecto del prior, estudiaré la dependencia de la diferencia entre las distribuciones posteriores obtenidas usando un prior objetivo y uno subjetivo, como función del número de datos observados.

Dependencia con la cantidad de los datos. Supongamos que la distribución *likelihood* dado un conjunto de datos $\{\mathcal{D}_i\}$ es

$$\mathcal{L}\left(\{\mathcal{D}_i\}\,|\,\theta, I\right) = \prod_{i=1}^{N} \mathcal{L}_i\left(\mathcal{D}_i\,|\,\theta, I\right),$$

entonces la log-posterior es simplemente

$$\log \pi_N\left(\theta\,|\{\mathcal{D}_i\}, I\right) = \sum_{i=1}^{N} \log \mathcal{L}_i\left(\mathcal{D}_i\,|\,\theta, I\right) + \log \pi_0\left(\theta\,|\,I\right) + K,$$

donde K es una constante. Es claro entonces que la distribución posterior escala con N, el número de datos. Esto tiene sentido intuitivamente hablando, porque es de esperarse que en el límite de $N \to \infty$ los datos sean lo suficientemente informativos a través de la verosimilitud como para anular la contribución de la distribución previa.

Dependencia calidad de los datos. Si los datos son de baja calidad, e. g. tienen incertidumbres típicas muy altas, la contribución de la verosimilitud de nuevo se ve comprometida, más aún en caso de una distribución previa sea no objetiva.

6. ¿Cómo elegir priors de manera apropiada?

Por supuesto la elección de la distribución de probabilidad prior depende del problema en particular que se quiera resolver, pero sobre todo del conocimiento que se posea en el espacio de la hipótesis. Por ejemplo, si nos sentimos muy seguros sobre nuestros prejuicios podríamos sentirnos tentados también a introducir un prior informativo, mientras que si por el contrario tenemos poco conocimiento sobre el problema lo mejor sería asumir una distribución previa objetiva. Pero ¿cómo representar lo que sabemos o ignoramos en forma de distribución de probabilidad de la forma más objetiva posible? Bueno, en el caso de una distribución previa objetiva es obvio: usamos el "principio de razón insuficiente", el cual supone que dado un conjunto exhaustivo y mutuamente excluyente de posibilidades, si no hay razones para pensar que una posibilidad es más probable que otra, entonces lo más justo es asignar la misma probabilidad a todas. Pero ¿qué pasa si en efecto poseemos razones suficientes para pensar que la distribución no es objetiva?

Información y entropía. La probabilidad de ocurrencia de un evento necesariamente implica que tenemos conocimiento de la naturaleza de este y el conocimiento es información. Pero ¿cuál es la relación entre la información y la probabilidad? Supongamos que sabemos la probabilidad $\Pr(\theta)$ de que un evento θ ocurra, ¿cuánta información obtenemos si observamos que el evento ocurre? Bueno, exploremos las reglas que debe cumplir la función de información $\mathrm{I}[\Pr(\theta)]$.

Rango: Por conveniencia se impondrá la restricción de que la información debe ser una cantidad tal que $\mathrm{I} \geqslant 0$.

Monotonía y continuidad: El dominio de I viene representado por el rango $\Pr = [0, 1]$. Si se tiene la certeza de que un evento ocurre, entonces la información obtenida por tras su observación es $\mathrm{I}(1) = 0$. Mientras menor es la probabilidad de observar un evento, la información obtenida tras su observación es mayor. Por lo tanto I debe ser una función decreciente en $\Pr$.

Aditividad: Si dos eventos independientes tienen probabilidad $\Pr(\theta_1)$ y $\Pr(\theta_2)$ de ocurrir, entonces su probabilidad conjunta es el producto de sus probabilidades individuales y la información provista por la observación de ambos eventos es $\mathrm{I}_{\theta_1,\theta_2} = \mathrm{I}(\Pr(\theta_1)) + \mathrm{I}(\Pr(\theta_2))$.

La última regla o axioma restringe considerablemente las posibilidades para I, de manera que la forma de representar dicha información de manera estándar es simplemente:

$$\mathrm{I}(\Pr) \equiv -\log \Pr.$$

Ahora supongamos que tenemos un conjunto exahustivo de eventos mutuamente excluyentes cuyas probabilidades previas forman una distribución de probabilidad $\Pr$, la información esperada debida a la observación de alguno de tales eventos es:

$$\mathrm{H}(\Pr) \equiv -\int \Pr \log \Pr \, \mathrm{d}\theta.$$

Esta cantidad es la denominada entropía de una distribución de probabilidad y es una medida de la impredictibilidad de los eventos que representa dicha distribución. Pero ¿qué propiedades posee la entropía? Uno puede demostrar fácilmente que la entropía tiene un máximo usando la desigualdad de Gibbs:

$$\int \Pr \, \log \left[\frac{1/(\theta_{\max} - \theta_{\min})}{\Pr} \right] \mathrm{d}\theta \leqslant 0,$$

con la igualdad si y solo si $\Pr = 1/(\theta_{\max} - \theta_{\min})$.

Método de Máxima Entropía. En el apartado anterior vimos que dada la probabilidad previa de ocurrencia de un evento se puede cuantificar la información que se obtiene tras su observación simplemente usando la probabilidad asignada. Además, la entropía es la información que se espera obtener dada una distribución de probabilidad. En este sentido, parece coherente que una forma justa de asignar la distribución previa, $\pi_0(\theta \mid I)$ consista en maximizar la entropía, esto es, maximizar la información que se obtiene una vez se ha hecho una observación.

Supongamos que tenemos información contrastable sobre una hipótesis θ y queremos asignar una distribución de probabilidad que cumpla con esa condición. El método de máxima entropía consiste en maximizar la función:

$$L(Pr;\theta,I) = \int Pr(\theta \mid I)\ \log\left[1/Pr(\theta \mid I)\right] d\theta - \lambda_0 \left[\int Pr(\theta \mid I)\ d\theta - 1\right] - \sum_{i=1}^{N} \lambda_i\, g_i(Pr;\theta,I),$$

donde el primer término es la entropía, el segundo es la condición de normalización y es una restricción necesaria y la sumatoria representa las restricciones impuestas por la información contrastable encapsulada en las funciones $g_i(Pr;\theta,I)$, normalizadas por los multiplicadores de Lagrange λ_i.

7. ¿Es el modelo adoptado el correcto?

La distribución posterior proporciona estimaciones de los parámetros y de sus incertidumbres, incluyendo la propagación de las incertidumbres observacionales, de manera consistente y confiable, siempre y cuando la distribución previa comprenda todas las posibilidades en el espacio de hipótesis. Sin embargo, por sí misma la distribución posterior no contiene información sobre si el modelo adoptado es bueno o malo.

Una forma de averiguar la plausibilidad del modelo adoptado consiste en muestrear la distribución posterior predictiva, esto es:

$$Pr(\{\mathcal{D}_i\}^{pre} \mid \{\mathcal{D}_i\}, I) = \int Pr(\{\mathcal{D}_i\}^{pre} \mid \theta, I) \times \pi_N(\theta \mid \{\mathcal{D}_i\}, I)\, d\theta, \qquad (3)$$

donde el primer factor dentro de la integral es la probabilidad de haber observado $\{\mathcal{D}_i\}^{pre}$ dada la hipótesis. La premisa es que si el modelo (o mejor dicho la verosimilitud) comprende el espacio de observaciones en su completitud, entonces una muestra tomada de la distribución $\mathcal{N}(\{\mu_i(\theta)\},\{\sigma_i\})$ debería ser indistinguible de las observaciones reales.

8. Ejemplo: Función Inicial de Masa

En este ejemplo vamos a reconstruir la FIM usando el método de máxima entropía. Como información verificable

9. Lecciones

A. Distribuciones conjugadas

B. Muestreo de la posterior

Existen varias formas de muestrear la posterior y la eficiencia de la técnica adoptada dependerá exclusivamente de la naturaleza del problema reflejada en la forma de la distribución posterior. Por ejemplo, si la distribución posterior se puede escribir analíticamente, no será necesario recurrir a métodos numéricos como los algoritmos de Monte Carlo Markov Chain (MCMC). Si en cambio la distribución posterior tiene una forma extraña que no se puede representar simplemente de

manera analítica, los métodos de muestreo aleatorio serán necesarios. Nótese que esto también aplica a las distribuciones prior, que también deben ser muestreadas aleatoriamente durante el proceso de obtención de la distribución posterior en estos casos (me refiero a los casos de muestreo numérico).

Comencemos por los casos simples, en los que la posterior es simplemente representada por una ecuación matemática. Definamos el concepto de distribución conjugada. Se llama distribución conjugada a cualquier distribución perteneciente a la misma familia. Por ejemplo, en el contexto de la estadística bayesiana, si la distribución prior y la distribución posterior son familias, se dice que el prior es el conjudado de la posterior y viceversa.