Inferencia en el marco de la Estadística Bayesiana

Alfredo Mejía-Narváez

2 de febrero de 2017

Resumen

La estadística bayesiana no es solo un conjunto de reglas diseñadas para la interpretación de los datos en nuevo conocimiento dada la experiencia previa, es una forma de pensar. De la gran mayoría de los libros de estadística aprendemos a responder a la pregunta equivocada o, al menos, una que realmente nos es la que nos interesa como científicos: 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 probabilidad de haber obervado los datos?. La estadística bayesiana no solo nos permite responder a la pregunta de interés científico, sino que nos permite darle una respuesta lo más objetivamente posible, mediante probabilidades. Entonces, la pregunta de interés es en general: dados la experiencia previa (nuestro grado de conocimiento del problema en ausencia de los datos) y los datos ¿qué podemos inferir para incrementar nuestro grado de conocimiento del problema?. En este seminario les presentaré el esquema general de la inferencia bayesiana, algunas de sus aplicaciones en la astrofísica y dos ejemplos: uno de juguete para ilustrar conceptos básicos y otro de la vida real.

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

$$P(\theta|I) + P(\bar{\theta}|I) = 1 \tag{1}$$

y

$$P(\theta_1, \theta_2|I) = P(\theta_1|\theta_2, I) \times P(\theta_2|I) \tag{2}$$

$P(\theta|I)$: 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, $\boldsymbol{\Theta}$.

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_2 y nuestro grado de conocimiento en θ_1 condicionado en θ_2, podemos admitir que sabemos nuestro grado de conocimiento de θ_1 y θ_2.

2. El teorema de Bayes

$P(\boldsymbol{\theta}|\{\mathcal{D}\}, I)$: es la distribución de probabilidad posterior y representa nuestro grado de conocimiento de la hipótesis a la luz de los datos, $\{\mathcal{D}\}$.

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

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

$$P(\boldsymbol{\theta}|\{\mathcal{D}\}, I) = \frac{P(\{\mathcal{D}\}|\boldsymbol{\theta}, I) \times P(\boldsymbol{\theta}|I)}{P(\{\mathcal{D}\}|I)}, \tag{3}$$

donde

$$P(\{\mathcal{D}\}|I) = \int_{\Theta} P(\boldsymbol{\theta}|\{\mathcal{D}\}, I)\mathrm{d}\boldsymbol{\theta}.$$

El poder del teorema de Bayes en sí yace en nuestra capacidad para asignar las distribuciones (densidades en realidad) de probabilidad en el lado derecho de la igualdad, $P(\{\mathcal{D}\}|\boldsymbol{\theta}, I)$ y $P(\boldsymbol{\theta}|I)$, en la Ec. (3). Mientras 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, $P(\boldsymbol{\theta}|\{\mathcal{D}\}, I)$. 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. Inferencia bayesiana: regresión lineal

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 datos. Supongamos que la distribución *likelihood* dado un conjunto de datos $\{\mathcal{D}\}$ es

$$P(\{\mathcal{D}\}|\boldsymbol{\theta}, I) = \prod_{i=1}^{N} P(\mathcal{D}_i|\boldsymbol{\theta}, I)$$

entonces la log-posterior es simplemente

$$\log P(\boldsymbol{\theta}|\{\mathcal{D}\}, I) = \sum_{i=1}^{N} \mathcal{L}(\mathcal{D}_i|\boldsymbol{\theta}, I) + \log P(\boldsymbol{\theta}|I) + K,$$

donde $\mathcal{L}(\{\mathcal{D}\}|\boldsymbol{\theta}, I) \equiv \log P(\{\mathcal{D}\}|\boldsymbol{\theta}, I)$ y 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 $N \to \infty$ la información que proveen los datos acerca de nuestra inferencia sea más relevante que la de la distribución prior.

Dependencia con la forma del prior. Es de esperarse que nuestra inferencia estadística sea fuertemente dependiente de nuestra elección de la distribución prior en el límite de pocos datos, aún cuando el prior sea objetivo. Sin embargo, siempre tendremos la oportunidad de introducir nuestros prejuicios (aún cuando $N \to \infty$) si el prior es lo suficientemente subjetivo o mejor dicho, informativo.

¿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. Por ejemplo, si el científico se siente muy confiado sobre sus prejuicios podría sentirse tentado también a introducir un prior informativo, mientras que si por el contrario tiene poco conocimiento del problema lo mejor sería asumir un prior objetivo. Como hemos visto anteriormente, independientemente del prior que se

6. Muestreo de la posterior

Existen varias formas de muestrar 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.

7. Marginalización

8. Chequeo del modelo

9. Ejemplo: X

10. Epílogo

A. Distribuciones conjugadas

B. Método de máxima entropía

Supongamos que tenemos información constrastable sobre una hipótesis $\boldsymbol{\theta}$ y queremos asignar una distribución de probabilidad que cumpla con esa condición y con la condición de que la distribución represente en sí una distribución de probabilidad, i.e. que se cumple la condición (1) ¿cuál es la forma más *justa* de asignar dicha distribución de probabilidad?