Inferencia en el marco de la Estadística Bayesiana

Alfredo Mejía-Narváez

7 de enero de 2018

Resumen

La estadística representa un conjunto de reglas diseñadas para expresar nuestro grado de conocimiento dados cierta evidencia y/o conocimiento previo y para avanzar dicho conocimiento a la luz de nuevas observaciones. Sin embargo, nuestra línea de razonamiento científico se ve forzada a desviarse un poco de este esquema, que es la forma natural, y la pregunta de interés científico real: ¿cuál es nuestro grado de conocimiento sobre un evento, dados un conjunto de observaciones y conocimiento previo o prejuicios? se transforma en otra cosa en la 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 probabilidad de haber observado los datos bajo cierta hipótesis?. La estadística Bayesiana presenta un esquema que permite responder naturalmente a la pregunta de interés científico y 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.

Introducción: estadística básica

El concepto de probabilidad ha venido a responder una pregunta fundamental para la 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:

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

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

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.

$\Pr(\theta \mid I)$: 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, $\boldsymbol{\Theta}$.

El teorema de Bayes

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

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

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

donde

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

$\Pr(\{\mathcal{D}_i\} \mid I)$: 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, $\mathcal{L}(\{\mathcal{D}_i\} \mid \theta, I)$ y $\pi_0(\theta \mid I)$, 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.*

Inferencia bayesiana

La inferencia bayesiana es la técnica que consiste en calcular la distribución de probabilidad posterior, $\pi_N(\theta \mid \{\mathcal{D}_i\}, I)$ 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.

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:

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

Una vez hemos llegado a esta realización, la pregunta: ¿cómo encontrar los parámetros m y b, que mejor ajustan a los datos? tiene una respuesta trivial: maximizar la verosimilitud o, equivalentemente, minimizar la función de mérito χ^2. En este caso, la respuesta a nuestra pregunta se reduce simplemente a: Y en el mejor de los casos a: En cuyo caso el ajuste es: Uno podría 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. Claramente en este caso, como la distribución previa no es informativa, la distribución posterior es igual a la verosimilitud. Pero ¿representan lo mismo? bueno, evidentemente el teorema de Bayes dice que no es así. Vamos a ver por qué desde el punto de vista de la información que se poseen. La función de verosimilitud representa una distribución de frecuencias en el espacio de observaciones dado un conjunto de parámetros que presumiblemente lo describen, lo que quiere decir que solo en el límite de infinitas observaciones estamos seguros de si esto es así o no.

Independientemente del método que se elija para obtener los parámetros que garantizan el mejor ajuste a los datos, siempre es bueno acompañarlos con incertidumbres.

Incertidumbres en MV. En el método de Máxima Verosimilitud, ¡la matriz de covarianza para los parámetros del mejor ajuste es independiente de la bondad del ajuste! De manera que la mejor estimación de la incertidumbre en este caso únicamente refleja el hecho de que las incertidumbres en las observaciones son Gaussianas y ni siquiera eso está garantizado en la mayoría de los casos. En general deben cumplirse tres condiciones para que la matriz de covarianza sea representativa de los errores en los parámetros estimados:

Las σ_i: las incertidumbres en las observaciones deben estar bien determinadas y representadas por σ_i, i. e., estas deben considerar no solo efectos instrumentales, sino también las variaciones intrínsecas del fenómeno observado.

Gaussianidad: el proceso (fenómeno y observación) debe ser completamente Gaussiano.

Modelo perfecto: el modelo asumido debe ser representativo de los datos, i. e., en ausencia de incertidumbres, la desviación del modelo debe ser nula.

En la práctica rara vez se cumplen estas condiciones. Además casi nunca se cuenta con una muestra de la verosimilitud, de manera que no es trivial calcular las incertidumbres en el resultado.

Incertidumbres en Inferencia Bayesiana. En la estadística Bayesiana, las incertidumbres en los parámetros depende de manera consistente del modelo asumido a través de la verosimilitud y del conocimiento inicial sobre los parámetros a través de la distribución previa. La inferencia Bayesiana viene acompañada de una muestra de la distribución posterior y, en ciertos casos muy especiales, de una forma analítica de esta, de manera que es relativamente sencillo calcular incertidumbres en los parámetros estimados.

En este sentido, si realmente no estamos seguros de que el modelo asumido (independientemente de si es lineal o no) realmente es un modelo generativo de todo el espacio de observaciones $\{y_i\}$, una forma de resolver este tipo de problemas es usando inferencia Bayesiana.

Bajo el esquema Bayesiano, la distribución previa no solo se usa para introducir conocimiento 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.

El efecto de la distribución previa

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 de la distribución previa.

Para cuantificar el efecto de la distribución previa, 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.

¿Cómo elegir $\pi_0\left(\theta\,|\,I\right)$?

Por supuesto la elección de la distribución de probabilidad previa 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 a asumir una distribución previa informativa, 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 plana?

Información y entropía. En la ciencia la observación de un evento casi siempre proporciona información que eventualmente nos permite actualizar nuestro grado de conocimiento sobre ese evento. Cuando uno diseña un experimiento, uno de los pasos es plantear una hipótesis que refleje nuestro conocimiento previo a la observación del evento. Estadísticamente hablando esta hipótesis es descrita por la probabilidad de ocurrencia de dicho evento. Pero ¿cuál es la relación entre la información adquirida tras la observación y la probabilidad previa de ocurrencia?

Supongamos que la probabilidad previa de un evento θ es $\Pr\left(\theta\,|\,I\right)$, ¿cuánta información obtenemos si observamos que el evento ocurre? Bueno, Shannon (quien diseñó la teoría de información) planteó un conjunto de reglas simples que debe cumplir la función de información $I[\Pr\left(\theta\,|\,I\right)]$ suponiendo que esta solo dependa de la probabilidad previa.

Dominio: La información depende de $\Pr(\theta \mid I)$ únicamente, de manera que su dominio es $[0, 1]$.

Rango: La información debe ser positiva ($I \geqslant 0$). Si se tiene la certeza de que un evento ocurre, entonces la información obtenida tras su observación es $I(1) = 0$. Mientras menor es la probabilidad previa de observar un evento, la información obtenida tras su observación es mayor.

Monotonía y continuidad: La información debe ser monótona y continua, esto es, pequeños cambios en $\Pr(\theta \mid I)$ deben reflejar pequeños cambios en I y *vice versa*.

Aditividad: Si dos eventos independientes tienen probabilidad $\Pr(\theta_A \mid I)$ y $\Pr(\theta_B \mid I)$ de ocurrir, entonces su probabilidad conjunta es el producto de sus probabilidades individuales y la información provista por la observación de ambos eventos es $I_{\theta_A,\theta_B} = I[\Pr(\theta_A \mid I)] + I[\Pr(\theta_B \mid I)]$.

La última regla restringe considerablemente las posibilidades para I, de manera que la forma de representar dicha información (demostrada por Shannon) es simplemente:

$$I[\Pr(\theta \mid I)] \equiv -\log \Pr(\theta \mid I).$$

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

$$H[\Pr(\theta \mid I)] \equiv -\int \Pr(\theta \mid I)\ \log \Pr(\theta \mid I)\, d\theta.$$

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? y, en particular ¿la entropía posee un máximo? (ya veremos por qué nos interesa precisamente el máximo) Uno puede demostrar fácilmente que la entropía tiene un máximo usando la desigualdad de Gibbs:

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

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

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 objetiva 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, I, sobre un conjunto exhaustivo de eventos mutuamente excluyentes $\boldsymbol{\theta}$ y queremos asignar una distribución de probabilidad que cumpla con esa condición y que a la vez garantice que posterior a la observación tendremos la máxima información posible. El método de máxima entropía consiste en optimizar la función:

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

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

Veamos un ejemplo interesante para ilustrar la utilidad de esta técnica para asignar distribuciones de probabilidad a partir de I.

EJEMPLO: FUNCIÓN INICIAL DE MASA. La Función Inicial de Masa (FIM) puede escribirse como una distribución de probabilidad en masa estelar, m, de la forma:

$$\Pr(m \mid I) = \gamma\, m^{-\alpha},$$

donde γ es una constante de normalización y α es el exponente. Vamos a tratar de recuperar la misma forma funcional de la FIM usando argumentos estadísticos y el método de la ME.

Supongamos que en la región alrededor de una protoestrella la densidad del gas que esta está acretando por la protoestrella i es ρ_i^g, si la tasa de acreción es:

$$R_i = \frac{m_i}{\tau_i^{\text{din}}},$$

con m_i la masa de la protoestrella y τ_i^{din} la escala de tiempo dinámica del entorno. Si la acreción ocurre en caída libre (i. e., la viscosidad del medio es despreciable) entonces:

$$R_i \propto (m_i\, r_i)^{3/2},$$

donde r_i es el radio de la protoestrella. La densidad de tasa de acreción es R_i/V_i, de manera que:

$$\rho_i^R \propto (m_i/r_i)^{3/2}.$$

Queremos encontrar entonces el conjunto $\{P_i\}$ que defina de probabilidad de que la estrella i tenga masa m_i cuando entre a la secuencia principal o lo que es igual, en el instante en que $R_i = 0$. Como las estrellas de secuencia principal siguen la relación de escala: $r_i = m_i^{0,8}$, entonces la ecuación de arriba queda simplemente como:

$$\rho_i^R \propto m_i^{0,2}.$$

Vamos a suponer además que el proceso de acreción estelar ocurre de manera que esta es una versión local de la Ley de Kennicutt-Schmidt (**?**) de formación estelar:

$$\rho_\Psi = \epsilon \rho_g^{3/2},$$

de manera que:

$$\sum_i P_i \, \log m_i \propto \log \rho_g + K,$$

Así, el Lagrangiano queda:

$$L = \sum_i P_i \, \log(1/P_i) - \lambda_0 \left[\sum_i P_i - 1\right] - \lambda_1 \left[A + B \sum_i P_i \log m_i - \log \rho_g\right],$$

donde A y B son constantes. Entonces:

$$\frac{\partial L}{\partial P_i} = -(1+\lambda_0) - B\lambda_1 \log m_i - \log P_i,$$

haciendo $\alpha \equiv B\lambda_1$ y $\gamma \equiv e^{-1-\lambda_0}$, concluimos que:

$$P_i = \gamma\, m_i^{-\alpha} \quad \Longrightarrow \quad \Pr(m \mid I) = \gamma\, m^{-\alpha}.$$

Hemos demostrado que la FIM puede explicarse por un proceso estocástico usando la Ley de formación estelar de KS.

¿QUÉ TAN BUENO ES EL MODELO?

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 cuantificar la plausibilidad del modelo adoptado consiste en muestrear la distribución posterior predictiva, esto es:

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

donde el primer factor dentro de la integral es la probabilidad de haber observado $\{\mathcal{D}_j\}^{\text{pre}}$ dada la hipótesis. La premisa es que si el modelo es descriptivo de las observaciones en su completitud, entonces una muestra tomada de la distribución $\Pr(\{\mathcal{D}_j\}^{\text{pre}} \mid \theta, I)$ debería ser indistinguible de las observaciones reales. Para verificar que esto es así, se ejecutan los siguientes pasos:

Conjunto $\{\theta_j\}^{\text{pre}}$. Necesitamos un conjunto actualizado (posterior) de parámetros en el espacio de hipótesis. Así que muestreamos la distribución posterior.

Conjunto $\{\mathcal{D}_i\}^{\text{pre}}$. Usando el conjunto θ^{pre} del paso anterior, se construye un conjunto de observaciones predichas, muestreando la distribución $\mathcal{P}(\mu \mid \theta^{\text{pre}}, I)$. Por supuesto $\Pr(\{\mathcal{D}_j\}^{\text{pre}} \mid \theta, I)$ y $\mathcal{P}$ deben ser consistentes: si la probabilidad del evento $\mathcal{D}_i$ se distribuye alrededor de un promedio μ_i con desviación estándar σ_i, entonces la elección sensible para $\mathcal{P}$ sería una distribución Gaussiana y la verosimilitud sería muy parecida a la del problema de ajustar una línea recta que vimos antes.