EST-25134 Aprendizaje Estadístico **ITAM**

Paradigmas Inferenciales

Prof.: Juan Carlos Martínez-Ovando *31 de enero de 2017*

1.1. Paradigma Frecuentista

1.1.1. Inferencia Estadistica y el Proceso de Aprendizaje

1.1.2. Prediccion]

1.2. Paradigma Bayesiano

La estadística es el estudio de fenómenos bajo un estado de conocimiento o información incompleto. Los fundamentos teóricos de lo que en la actualidad se conoce como estadística Bayesiana tienen su origen con la publicación de un artículo del Reverendo Thomas Bayes en 1773, dos años después de su muerte. En ese trabajo, Thomas Bayes resolvió un problema de información inversa planteado por Bernoulli, que consiste en obtener información sobre réplicas independientes de variables aleatorias Bernoulli. Una década después, Laplace retomó las ideas de Bayes y desarrolló con mayor claridad lo que en la actualidad se conoce como el paradigma Bayesiano de inferencia.

Denotemos por $\mathscr{H}_1, \mathscr{H}_2, ..., \mathscr{H}_p$ a una colección de proposiciones o hipótesis excluyentes y exhaustivas, y supongamos que deseamos realizar inferencias sobre éstas con base en un nivel de información denotado por $\mathscr{I}$, el cual resume nuestra percepción e información inicial respecto a estas hipótesis. Nuestro estado de información lo expresamos a través de una medida de probabilidad definida sobre el espacio de las hipótesis o proposiciones en cuestión, condicional en nuestro estado de información, que denotamos por $P(\mathscr{H}_i|\mathscr{I})$, y es tal que $P(\mathscr{H}_1|\mathscr{I}) + \cdots + P(\mathscr{H}_p|\mathscr{I}) = 1$. Nuestro aprendizaje respecto a las hipótesis consiste en la actualización del conocimiento mediante la incorporación de nueva información relevante, que denotamos

por $\mathscr{D}$. Por simetría tenemos la siguiente relación

$$P(\mathscr{D}|\mathscr{I})\,P(\mathscr{H}_i|\mathscr{D},\mathscr{I}) = P(\mathscr{H}_i|\mathscr{I})\,P(\mathscr{D}|\mathscr{H}_i,\mathscr{I}), \tag{1.1}$$

para $i = 1,...,p$. Si $P(\mathscr{D}|\mathscr{I}) > 0$, i.e. la información relevante proporcionada por el entorno real es plausible, entonces nuestro estado de información actualizado es de la forma

$$P(\mathscr{H}_i|\mathscr{D},\mathscr{I}) = P(\mathscr{H}_i|\mathscr{I})\,\frac{P(\mathscr{D}|\mathscr{H}_i,\mathscr{I})}{P(\mathscr{D}|\mathscr{I})}. \tag{1.2}$$

La relación (1.2) es la representación matemática del proceso de aprendizaje y es conocida como el Teorema de Bayes, aún cuando Bayes no haya sido quien lo enunció formalmente. Esta relación muestra cómo la probabilidad inicial o *a priori* respecto a las hipótesis, $P(\mathscr{H}_i|\mathscr{I})$, es actualizada a la probabilidad final o *a posteriori,* $P(\mathscr{H}_i|\mathscr{D},\mathscr{I})$, como resultado de la incorporación de nueva información $\mathscr{D}$. El Teorema de Bayes puede ser aplicado repetidamente conforme nueva información $\mathscr{D}_1,\mathscr{D}_2,...$ es obtenida, en cuyo caso la distribución final se convierte en la nueva información inicial para el caso siguiente, de forma que en cualquier instante la *plausibilidad* de la hipótesis $\mathscr{H}_i$ dependerá de la evidencia total disponible. De esta forma, captura la naturaleza secuencial del proceso de aprendizaje general que usualmente efectuamos en nuestra vida cotidiana.

Durante los años subsecuentes del siglo XVIII y del siguiente, el paradigma Bayesiano se encontró en un estado inerte debido a que esta teoría carecía de sustentabilidad teórica respecto al enfoque frecuentista de inferencia. No fue sino hasta el segundo tercio del siglo pasado en que Harold Jeffreys y Bruno de Finetti desarrollaron y formalizaron la teoría que actualmente se encuentra vigente. Ambos fueron defensores del paradigma Bayesiano, aunque tenían visiones distintas respecto a la conceptualización e interpretación de la probabilidad. Por un lado, Jeffreys defendió una postura *objetiva* sobre el tema, y por otro lado de Finetti propuso y formalizó una visión enteramente *subjetiva* donde se entiende que la probabilidad mide el grado de creencia respecto al fenómeno de interés del individuo quien la expresa. En este sentido, la información inicial en la relación del (1.2) es necesariamente subjetiva. La interpretación objetiva e impositiva de Jeffreys, desarrollada de manera más flexible por Richard Cox, se basa en un principio de consistencia, el cual enuncia que dos individuos con el mismo nivel de información deban necesariamente reportar la misma apreciación inicial respecto a su incertidumbre de manera que las conclusiones que éstos generan necesariamente deben de ser completamente compatibles. Con el enfoque de de Finetti esta regla no necesariamente debe de cumplirse.

El enfoque Bayesiano ha evolucionado de manera sorprendente durante los años subsecuentes. Su uso nos provee de una herramienta útil de inferencia y sobre todo de predicción, que en general puede considerarse como el problema central del análisis estadístico. Una revisión detallada respecto a la evolución del paradigma Bayesiano, y en general del proceso de inferencia estadística, la podemos encontrar en **?** y **?**, entre otros. En las siguientes subsecciones describiremos los principios fundamentales de inferencia estadística Bayesiana y predicción.

1.2.1. Inferencia Estadística y el Proceso de Aprendizaje

Supongamos que una variable aleatoria de interés, denotada por Y, tiene una distribución de probabilidad en la familia $\mathscr{P} = \{p(y|\theta) : \theta \in \Theta\}$, donde θ es un parámetro que indiza la función de probabilidad de la variable aleatoria Y, y Θ es un espacio parametral. Desde el enfoque Bayesiano el desconocimiento sobre el valor del parámetro de interés θ es manifestado mediante la asignación de una medida de probabilidad, digamos $\pi(\theta)$, que representa nuestro nivel de información sobre el verdadero valor de éste. Denotemos por y a un conjunto de realizaciones observables de la variable Y, y denotemos a la distribución de probabilidad conjunta de y y θ por $p(y,\theta)$. Entonces, por las leyes básicas de probabilidad, se cumplen las siguientes relaciones

$$p(y,\theta) = p(y|\theta)\pi(\theta) = \pi(\theta|y)p(y) \tag{1.3}$$

donde $p(y|\theta)$ es la función de probabilidad de la v.a. y condicional en θ; y $p(y)$ y $\pi(\theta)$ son las funciones de densidad marginales de y y θ respectivamente.

De las ecuaciones anteriores es posible deducir que

$$\begin{aligned} \pi(\theta|y) &= \frac{p(y|\theta)\pi(\theta)}{p(y)} \\ &\propto p(y|\theta)\pi(\theta) \\ &\propto \textit{verosimilitud} \times \textit{inicial} \end{aligned} \tag{1.4}$$

donde $\propto$ denota proporcionalidad en términos de θ; $\pi(\theta|y)$ es conocida como la distribución *final* (o *a posteriori*) del parámetro θ, condicional a la información muestral y**;** $\pi(\theta)$ es la distribución *inicial* (o *a priori*) asignada al parámetro θ; y $p(y|\theta)$ es la función de verosimilitud, vista como función de θ.

La distribución inicial del parámetro $\pi(\theta)$ cuantifica nuestro estado de información respecto al valor desconocido del parámetro θ. Este conocimiento lo actualizamos, mediante la aplicación del Teorema de Bayes,

con la incorporación de información adicional relevante, por ejemplo una muestra aleatoria observada y de la variable de interés, que proporcione evidencia sobre el verdadero parámetro θ. De esta forma, nuestro conocimiento actualizado sobre el parámetro θ es resumido en la distribución final, o *a posteriori*, $\pi(\theta|y)$.

Consideremos ahora que el parámetro está particionado como $\theta = (\theta_1, \theta_2)$, y que sólo un subconjunto de éste, digamos θ_1, es de interés inferencial; en este caso θ_2 es conocido como parámetro de ruido. Dada una muestra y**,** es de interés encontrar la distribución marginal final de θ_1, sin prestar atención al valor de θ_2. De esta forma la inferencia que se realice sobre θ_1 deberá basarse en la distribución final de θ_1 condicional en y, la cual obtenemos con el siguiente proceso de marginalización

$$\begin{aligned} \pi(\theta_1|y) &= \int \pi(\theta_1, \theta_2|y)\, d\theta_2 \\ &= \int \pi(\theta_1|\theta_2, y)\, \pi(\theta_2|y)\, d\theta_2 \\ &= \mathbb{E}_{\Theta_2|y}\{\pi(\theta_1|\theta_2, y)\} \end{aligned}$$

donde Θ_2 es el espacio parametral de θ_2; $\pi(\theta_1|\theta_2, y)$ es la distribución final condicional de θ_1 dado θ_2 y y**;** y $\pi(\theta_1|y)$ es la distribución final marginal de θ_1 dado y.

La aplicación del paradigma Bayesiano nos permite establecer un procedimiento secuencial de actualización de la información sobre el parámetro de interés θ. Supongamos que y_1 es una realización de la v.a. Y. Aplicando el Teorema de Bayes (1.4) la distribución final de θ dado y_1 es $\pi(\theta|y_1)$. Si posteriormente se tiene acceso a otra realización de Y, denotada por y_2, entonces la distribución final de θ dado y_1 y y_2, que resume nuestro conocimiento sobre θ actualizado por y_1 y y_2, puede expresarse como

$$\pi(\theta|y_1, y_2) \propto p(y_2|\theta, y_1)\pi(\theta|y_1). \tag{1.5}$$

De la ecuación (1.5) se puede establecer un procedimiento de aprendizaje secuencial, si se considera a $\pi(\theta|y_1)$ como la nueva distribución inicial para θ, antes de observar y_2.

1.2.2. Predicción

Uno de los objetivos centrales del análisis estadístico es el de predecir valores futuros de una variable aleatoria de interés Y condicional en la información histórica observada de la misma variable, y, y posiblemente bajo las consideraciones de algunos otros elementos o factores adicionales relevantes. Usando el enfoque Bayesiano, los resultados siguiendo este objetivo, se resumen a través de una distribución de probabilidad

definida sobre el espacio de las variables futuras, i.e. toda la información relevante sobre la variable futura, denotada por Y_f, estará resumida en $p(y_f|y)$, cuyo cálculo se obtiene de manera directa usando en Teorema de Bayes y un proceso simple de marginalización. De la ecuación (1.3) tenemos que

$$p(y_f, \theta|y) = p(y_f|\theta, y)\pi(\theta|y),$$

donde $p(y_f|\theta, y)$ es la densidad de la variable Y_f condicional en θ y y; y $\pi(\theta|y)$ es la distribución final de θ dado y. De esta forma, la distribución predictiva final la podemos calcular como

$$\begin{aligned} p(y_f|y) &= \int p(y_f, \theta|y)d\theta \\ &= \int p(y_f|\theta, y)\pi(\theta|y)d\theta \\ &= \mathbb{E}_{\Theta|y}\left\{p(y_f|\theta, y)\right\}. \end{aligned}$$

En el caso de variables aleatorias intercambiables, i.e. cuando $Y_1, ..., Y_n$ son condicionalmente independientes dado el parámetro θ, el cálculo de la densidad final de y_f se obtiene a través de

$$p(y_f|y) = \int p(y_f|\theta)\pi(\theta|y)d\theta,$$

en vista de la independencia condicional de y_f y y dado θ. El problema de inferencia y predicción es en esencia un problema de decisión estadística bajo un ambiente de incertidumbre. En la siguiente sección describiremos brevemente los elementos que conforman un problema de decisión y la solución Bayesiana óptima en el caso de inferencia o predicción puntual e inferencia y predicción general.

1.3. Elementos de la Teoría de Decisión

El problema estadístico de inferencia y predicción es básicamente un problema de decisión en un ambiente de incertidumbre: Un problema de decisión general está compuesto por un espacio de *estados de la naturaleza*, que denotaremos por Ω. En este espacio está definido el elemento sobre el cual reside nuestra incertidumbre y sobre el cual no tenemos ningún control. El espacio donde tenemos un control directo define nuestras diferentes alternativas o cursos de acción respecto al fenómeno o variable que nos interesa, y básicamente representa nuestras opciones disponibles en la búsqueda de un objetivo. Este espacio lo denotamos por $\mathscr{A}$. Cada trayectoria de decisión está compuesta por la pareja (a, ω) donde $a \in \mathscr{A}$ es la acción

o postura que hemos asumido respecto a la cantidad que nos interesa $\omega \in \Omega$, sobre la cual, como ya mencionamos, carecemos de control. Desde luego, las acciones tomadas nos conducirán a obtener diferentes resultados, que son desconocidos, y debemos de definir una escala de preferencias de manera que nuestras acciones sean consistentes y coherentes. Esta escala de preferencias sobre todas las posibles trayectorias de decisión la podemos definir a través una función de utilidad (o pérdida según sea el caso), inducida por nuestra relación de preferencia particular, y denotada por $u : \mathscr{A} \times \Omega \to \Re_+$, (o $l = -u$ en el caso de una función de pérdida).

Con el enfoque Bayesiano toda la información sobre el estado de la naturaleza, ω, está resumida en una medida de probabilidad $p(\cdot)$ condicional en toda la información relevante disponible al momento de la toma de decisiones. La solución Bayesiana óptima consiste en elegir la acción $a^* \in \mathscr{A}$ que maximice (minimice) la utilidad (pérdida) esperada (**?**, Capítulo 2).

En las siguientes subsecciones describiremos brevemente cómo se pueden obtener soluciones Bayesianas óptimas al problema de inferencia y predicción usando esta herramienta de toma de decisiones.

1.3.1. Estimación y Predicción Puntual

En algunas circunstancias, cuando la distribución de una v.a. Y está caracterizada por un valor parametral θ desconocido, es de interés encontrar un valor específico de θ, digamos θ^*, que describa convenientemente la distribución de probabilidad de la v.a. Y.

Claramente este es un problema de toma de decisiones en un ambiente de incertidumbre, donde el espacio de *estados de la naturaleza* y el espacio de acciones coinciden con el espacio parametral Θ. En este caso asignamos una medida de penalización a la acción de elegir un valor específico $\theta^* \in \Theta$ respecto al verdadero valor de $\theta \in \Theta$ (**?**). Por su naturaleza, esta función es conocida como *función de pérdida* y es denotada por $l(\theta, \theta^*)$, que al ser función de $\theta \in \Theta$ es una variable aleatoria.

Como ya mencionamos, la estrategia Bayesiana óptima consiste en elegir el valor θ^* que minimice la función de pérdida esperada respecto a la distribución final de θ dados los datos y, i.e. elegiremos $\theta^* \in \Theta$ tal que

$$\theta^* = \arg\min \left\{ \mathbb{E}_{\Theta|y} \left[l(\theta, \theta^*) | y \right] \right\}.$$

Usando, por ejemplo, la función de pérdida cuadrática, se tiene que el estimador puntual Bayesiano θ^* es

la media de la distribución final de θ, i.e. $\theta^* = \mathbb{E}(\theta|y)$. Otros estimadores puntuales, como la mediana y la moda de la distribución final de θ, pueden obtenerse como una solución alternativa si se utilizan ciertas funciones de pérdida (**?**).

Si nuestro interés reside en pronosticar un valor de la v.a. Y, con base en observaciones previas de la misma, entonces los estimadores de pronóstico Bayesiano los construiremos bajo el criterio anterior en términos de la distribución *predictiva final* de Y.

1.3.2. Inferencia y Predicción General

Supongamos que el interés del análisis estadístico es el de inferir respecto a un *estado de la naturaleza*, denotado por $\omega \in \Omega$, que se rige de manera aleatoria y sobre el cual nuestra información es limitada e inclusive en algunos casos faltantes. Las decisiones en este caso consisten en proporcionar alguna aseveración estadística respecto al valor incierto de ω, que desde un enfoque Bayesiano es resumida en una medida de probabilidad. Desde luego estas aseveraciones estarán condicionadas en la información relevante disponible al momento de la toma de decisiones, la cual en este caso denotamos por D, y que en términos generales está constituida por un conjunto de datos observados relacionados con el problema. En este caso el espacio de acciones estará definido por $\mathscr{A} = \{p_i(\cdot|D) : i \in I\}$, donde $p_i(\cdot|D)$ es una medida de probabilidad definida en Ω, para $i \in I$ con I un conjunto índice. Así, el conjunto de todas las posibles trayectorias del problema de decisión estarán denotadas por el conjunto $\mathscr{C} = \{c_i : i \in I\}$, donde $c_i = \{p_i(\cdot|D), \omega\}$ para todo $\omega \in \Omega$. La especificación de un problema de decisión general requiere establecer una relación de preferencia que cuantifique la consecuencia de decidir por el modelo $p_i(\cdot|D)$ cuando el estado de la naturaleza es ω.

La relación de preferencias se define en términos de una función de puntaje (**?**, definición 3.15) $u : \mathscr{A} \times \Omega \to \Re$. Así, la solución Bayesiana óptima consiste en elegir la distribución (o densidad) de la clase $\mathscr{A}$ que maximice en I la utilidad esperada

$$\overline{u}(p_i(\cdot|D)) = \int u(p_i(\cdot|D), \omega) p(\omega|D) d\omega, \tag{1.6}$$

donde $p(\omega|D)$ es la "verdadera" densidad de ω condicional en los datos observados D.

Se dice que una función de puntaje es *propia* si la utilidad esperada máxima se obtiene cuando $\sup_{i \in I} \overline{u}(p_i(\cdot|D)) = \overline{u}(p(\cdot|D))$, i.e. cuando la opción óptima es la "verdadera" densidad (distribución) para ω, y la función de

puntaje es *local* si $u(p_i(\cdot|D),\omega) = u(p_i(\omega|D))$ para todo $\omega \in \Omega$, i.e. si depende sólo del valor de la densidad (distribución) evaluada en ω.

? demostró que si una función de puntaje es propia y local, entonces debe ser de la forma

$$u\left(p_i(\cdot|D),\omega\right) = A\log p_i(\omega|D) + B(\omega), \tag{1.7}$$

para todo $\omega \in \Omega$, con $A > 0$ una constante real y $B(\cdot)$ una función integrable respecto a $p(\cdot|D)$. La función (1.7) es conocida como *función de puntaje logarítmico*.

1.4. Integración de Monte Carlo

Como vimos en las secciones anteriores, resolver un problema estadístico con el enfoque Bayesiano consiste operativamente en resolver integrales. En la práctica, muchas de estas integrales pueden ser difíciles de trabajar analíticamente. A través de la historia se han propuesto diferentes métodos para resolver algunos problemas de integración con estas características, algunos de los cuales consisten en aproximaciones numéricas deterministas o analíticas a la integral de interés. Las aproximaciones analíticas se basan en la aproximación de Laplace y resultan particularmente útiles para el caso de modelos cuya distribución pertenece a la familia exponencial dedistribuciones. En esta sección describiremos el método de Monte Carlo, que sirve para aproximar integrales complejas mediante técnicas de simulación estocástica. Para efectos prácticos supongamos que deseamos resolver una integral de la forma $\int g(\theta)\pi(\theta)d\theta$, donde $\theta \in \Theta \subset \Re^p$ es una variable aleatoria, $\pi(\cdot)$ es la densidad de θ, y $g(\cdot)$ es una función real conocida e integrable respecto a π. En el enfoque Bayesiano $\pi(\theta)$ estará condicionada en la información relevante disponible al momento del análisis, denotada por D, que por simplicidad en la notación es omitida en el transcurso de esta sección. Los resultados de esta sección son aplicables en los casos en que θ represente algunos parámetros asociados a un modelo, o cuando represente variables aleatorias observables.

El método de Monte Carlo se basa en el supuesto que seamos capaces de generar una muestra de tamaño N, $\{\theta^{(1)},...,\theta^{(N)}\}$, de la distribución $\pi(\theta)$[1]. Usando esta muestra podemos aproximar el valor de la integral de interés, la cual podemos interpretar como el valor esperado de g,

$$\mathbb{E}_\pi\left[g(\theta)\right] = \int g(\theta)\pi(\theta)d\theta, \tag{1.8}$$

[1] Por simplicidad π denotará a la distribución de θ y a su densidad de manera indistinta.

mediante el promedio empírico

$$\widehat{\mathbb{E}}_{\pi}[g(\theta)] = \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}). \tag{1.9}$$

El estimador (1.9), conocido como el estimador de Monte Carlo de (1.8), es un estimador insesgado y converge casi seguramente al valor de la integral de interés. Cuando la esperanza de $g^2(\cdot)$ es finita respecto a $\pi(\cdot)$, la convergencia de (1.9) puede medirse en términos de su varianza teórica

$$var\left[\widehat{\mathbb{E}}_{\pi}[g(\theta)]\right] = \frac{1}{N}\int [g(\theta) - \mathbb{E}[g(\theta)]]^2 \pi(\theta) d\theta, \tag{1.10}$$

la cual puede estimarse usando la misma muestra mediante su contraparte muestral

$$\widehat{var}\left[\widehat{\mathbb{E}}_{\pi}[g(\theta)]\right] = \frac{1}{N^2}\sum_{i=1}^{N}\left[g(\theta^{(i)}) - \widehat{\mathbb{E}}_{\pi}[g(\theta)]\right]^2. \tag{1.11}$$

En un contexto de inferencia Bayesiana generalmente conocemos la densidad $\pi(\cdot)$ salvo por una constante de normalización, que usualmente es difícil de calcular, y de hecho nos remonta al problema inicial de resolver una integral como (1.8) con $g(\cdot)$ igual a la función constante unitaria. En este caso es difícil generar datos de la distribución $\pi(\cdot)$ directamente, y por ende es difícil aplicar el método de Monte Carlo. A través de la historia se han propuesto diferentes alternativas para generar datos de densidades conocidas salvo por una constante de normalización. Algunos de éstos los describiremos brevemente en las siguientes subsecciones.

1.4.1. Muestreo por Importancia

El muestreo por importancia consiste en suponer que tenemos acceso a una densidad $p(\cdot)$ "semejante" a la densidad de interés $\pi(\cdot)$, conocida como *función de densidad de importancia*, de la cual es relativamente simple generar datos muestrales. La idea consiste en utilizar estos datos para aproximar integrales de la forma (1.8) usando el método de Monte Carlo.

La base central de este método consiste en suponer que el soporte de $p(\cdot)$ contiene al soporte de la densidad de interés $\pi(\cdot)$, en cuyo caso (1.8) puede ser re-expresada como

$$\int g(\theta)\frac{\pi(\theta)}{p(\theta)} p(\theta) d\theta = \mathbb{E}_p[g(\theta)w(\theta)],$$

donde $w(\theta) = \pi(\theta)/p(\theta)$. De esta forma, si podemos generar una muestra de la densidad $p(\cdot)$, de tamaño N, $\{\theta^{(i)} : i = 1,...,N\}$, entonces podemos aproximar (1.8) mediante

$$\widehat{\mathbb{E}}_{\pi}[g(\theta)] = \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)})w(\theta^{(i)}). \tag{1.12}$$

Cuando $\pi(\cdot)$ es conocida salvo por una constante de normalización, la aproximación (1.12) no puede ser usada directamente, sin embargo podemos expresar (1.8) como el cociente de dos esperanzas

$$\mathbb{E}_\pi[g(\theta)] = \frac{\int g(\theta)\frac{\pi(\theta)}{p(\theta)}p(\theta)d\theta}{\int \frac{\pi(\theta)}{p(\theta)}p(\theta)d\theta}, \tag{1.13}$$

en cuyo caso $\mathbb{E}_\pi[g(\theta)]$ puede aproximarse como el cociente de dos aproximaciones de Monte Carlo como

$$\widehat{\mathbb{E}}_\pi[g(\theta)] = \sum_{i=1}^{N} g(\theta^{(i)})\widetilde{w}(\theta^{(i)}),$$

donde $\widetilde{w}(\theta^{(i)}) = w(\theta^{(i)})/\sum_{j=1}^{M} w(\theta^{(j)})$ son los pesos asociados a cada dato $\theta^{(i)}$, con $w(\theta^{(i)})$ definida como antes, para $i = 1,...,M$. En este caso se hace evidente que no necesitamos la constante de normalización de la densidad de interés $\pi(\cdot)$.

La convergencia de (1.12) a (1.8) se garantiza si elegimos $p(\cdot)$ de manera que su soporte contenga al soporte de la densidad de interés $\pi(\cdot)$. Una consideración adicional para tener una convergencia más rápida es que el cociente $\pi(\theta)/p(\theta)$ esté acotado para todos los valores de θ, y que adicionalmente las colas de $p(\cdot)$ sean más pesadas respecto a las colas de $\pi(\cdot)$. Una descripción detallada de este método se encuentra en **?** describe algunas condiciones adicionales.

1.4.2. Muestreo-Remuestreo por Importancia

Este método extiende de manera natural las aproximaciones del método de muestreo por importancia.

El algoritmo funciona en dos etapas. En la primera etapa suponemos que tenemos una muestra de tamaño N de una densidad de importancia $p(\cdot)$, al igual que en la subsección anterior, i.e. tenemos una muestra $\{\theta^{(i)} : i = 1,...,N\}$ de la densidad $p(\theta)$, donde cada $\theta^{(i)}$ tiene un peso asociado $\widetilde{w}(\theta^{(i)})$, definido como en la subsección anterior. La segunda etapa del algoritmo consiste en generar N muestras con reemplazo de los valores $\{\theta^{(i)} : i = 1,...,N\}$ de acuerdo a sus correspondientes pesos $\{\widetilde{w}(\theta^{(i)}) : i = 1,...,N\}$. De esta forma podemos aproximar (1.8) por

$$\widehat{\mathbb{E}}[g(\theta)] = \sum_{j=1}^{N} g(\widetilde{\theta}^{(j)})\widetilde{w}(\widetilde{\theta}^{(j)}), \tag{1.14}$$

donde $\{\widetilde{\theta}^{(j)} : j = 1,...,N\}$ son una muestra de la variable discreta $\widetilde{\Theta} = \{\theta^{(i)} : i = 1,...,N\}$, donde cada $\theta^{(i)}$ tiene asociada una masa de probabilidad $\widetilde{w}(\theta^{(i)})$.

Más aún, con este procedimiento podemos aproximar características de $\pi(\cdot)$ que no pueden ser expresadas en forma de esperanza, como cuantiles e intervalos de credibilidad, ya que la distribución que asigna una masa $\widetilde{w}(\theta^{(i)})$ a $\theta^{(i)}$ en $\widetilde{\Theta}$ tiende en distribución a $\pi(\theta)$ cuando $N \to \infty$.

Este procedimiento es flexible y rico, en el sentido que podemos obtener muestras aproximadas que nos permiten reconstruir a $\pi(\cdot)$, por ejemplo a través de histogramas o aproximaciones por *kernel* (vea el apéndice **??**). Además permite implementar el Teorema de Bayes de manera directa, donde $\pi(\theta|y) \propto p(y|\theta)\pi(y)$. Si podemos generar una muestra aleatoria $\{\theta^{(i)} : i = 1,...,N\}$ de $\pi(\theta)$, podemos actualizarla a través de la verosimilitud para obtener una muestra $\{\widetilde{\theta}^{(j)}\}$ de tamaño N, que se distribuya aproximadamente como $\pi(\theta|y)$, remuestreando de $\{\theta^{(i)} : i = 1,...,M\}$, donde cada $\theta^{(i)}$ tiene asociado una masa de probabilidad definida como $\widetilde{w}(\theta^{(j)}) = p(y|\theta^{(j)})/\sum_{i=1}^{N} p(y|\theta^{(i)})$ para $j = 1,...,N$.

1.4.3. Monte Carlo vía Cadenas de Markov

Otro método importante para generar muestras de una distribución de probabilidad $\pi(\cdot)$ de interés, es construyendo una cadena de Markov cuya distribución invariante sea nuestra distribución objetivo $\pi(\cdot)$.

Supongamos que podemos construir una cadena de Markov homogénea $(\theta^{(n)})_{n\geq 1}$ en tiempo discreto, con un espacio de estados $\Theta \subset \Re^p$. En un esquema general, esta cadena de Markov está determinada mediante una función $K : \Theta \times \mathscr{B}(\Theta) \to [0,1]$ de transición de estados, conocida como *kernel de transición*, donde $\mathscr{B}(\Theta)$ es el σ-álgebra de Borel inducido por Θ. En el caso que Θ sea continuo, el *kernel* de transición denota a la densidad condicional de transición, $K(\theta,\theta')$, tal que $P(\Theta \in A|\theta) = \int_A K(\theta, d\theta')$. Cuando el espacio de estado Θ es discreto, el *kernel* de transición denota la probabilidad de transición $K(\theta,\theta') = P(\Theta^{(k+1)} = \theta'|\Theta^{(k)} = \theta)$ para todo θ y $\theta' \in \Theta$ entre las iteraciones k y $k+1$.

La idea central del método de Monte Carlo vía Cadenas de Markov (MCCM) es que la cadena de transición definida por un *kernel* de transición $K(\cdot,\cdot)$ tenga a $\pi(\cdot)$, la distribución de interés, como distribución *invariante*[2]. Este enfoque de análisis de cadenas de Markov es inverso al enfoque tradicional, ya que debemos construir una cadena partiendo de la distribución invariante, en lugar de construir una cadena con un *kernel* arbitrario y verificar si cumple con las condiciones de estabilidad. Si somos capaces de definir un *kernel*

[2] π es la densidad invariante de la cadena de Markov definida por el *kernel* $K(\cdot,\cdot)$ si $\theta^{(k)} \sim \pi$ implica que $\theta^{(k+1)} \sim \pi$, i.e. $\lim_{k\to\infty} K^k(\theta,A) = \pi(A)$, para todo $A \in \mathscr{B}(\Theta)$.

de transición que satisfaga la condición de balance $K(\theta,\theta')\pi(\theta) = K(\theta',\theta)\pi(\theta')$ para todo θ y $\theta' \in \Theta$, entonces tenemos que la cadena de Markov construida con este *kernel* tiene como densidad invariante a $\pi(\cdot)$ (**?**, Teorema 6.2.2). Si la cadena es *irreducible*[3] y *aperiódica*[4], entonces (**?**)

- $\theta^{(k)} \xrightarrow{d} \theta \backsim \pi$, y
- $\frac{1}{N}\sum_{k=1}^{N} g(\theta^{(k)}) \to \int g(\theta)\pi(\theta)d\theta$, casi seguramente cuando $N \to \infty$.

En las siguientes subsecciones describiremos diferentes métodos para construir cadenas de Markov con estas características.

1.4.4. Algoritmo de Metropolis-Hastings (M-H)

Para una cadena de Markov $(\theta^{(k)})_{k\geq 1}$, elegimos una familia de densidades $q(\theta,\theta')$ parametrizadas por θ, i.e. para un valor de θ fijo $q(\theta,\cdot)$ es una densidad con el mismo soporte que la densidad objetivo $\pi(\cdot)$. La elección de esta familia es arbitraria con el requisito que la cadena definida por la densidad de transición $P(\Theta^{(k+1)} = \theta' | \Theta^{(k)} = \theta) = q(\theta,\theta')$ sea irreducible, y que satisfaga la condición de balance.

El algoritmo funciona de la siguiente manera. Dado un estado actual de la cadena, digamos $\Theta^{(k)} = \theta^{(k)}$, un valor θ' es propuesto para el estado $\Theta^{(k+1)}$ con base en la densidad de transición $q(\theta^{(k)},\theta')$, y es aceptado con una probabilidad

$$\alpha(\theta^{(k)},\theta') = \text{mín}\left(1, \frac{q(\theta',\theta^{(k)})\pi(\theta')}{q(\theta^{(k)},\theta')\pi(\theta^{(k)})}\right), \tag{1.15}$$

i.e. con probabilidad α el valor de la cadena en la iteración $k+1$ es $\theta^{(k+1)} = \theta'$, de lo contrario $\theta^{(k+1)} = \theta^{(k)}$. Este esquema de muestreo define una cadena de Markov con un *kernel* de transición de la iteración k a $k+1$ dada por

$$K(\theta^{(k)},d\theta') = q(\theta^{(k)},d\theta')\alpha(\theta^{(k)},\theta') + \left(1 - \int \alpha(\theta^{(k)},\theta')q(\theta^{(k)},d\theta')\right)\delta_{\theta^{(k)}}(d\theta').$$

[3]Una cadena de Markov es *irreducible* (π-irreducible) si para todo $\theta \in E \in \mathscr{B}(\Theta)$ tal que $\pi(E) > 0$ se tiene que para todo $A \in \mathscr{B}(\Theta)$ con $\pi(A) > 0$ existe algún entero n tal que $K^n(\theta,A) > 0$, i.e. si existe la libertad de que la cadena se mueva sobre todo el espacio de estados.

[4]Una cadena de Markov es *aperiódica* si no existe una partición $\{E_0,...,E_{d-1}\}$ del espacio de estados Θ tal que $K(\theta,E_j) = 1$ para todo $\theta \in E_{j-1}$, i.e. no existe una trayectoria determinista de visitas a subconjuntos de Θ.

? demostró que una cadena de Markov construida de esta forma es reversible y aperiódica, con lo cual se tiene que $\pi(\cdot)$ es su correspondiente distribución estacionaria. Este algoritmo es particularmente útil en el contexto de inferencia Bayesiana, donde en algunas ocasiones $\pi(\theta)$ es conocida salvo su constante de normalización, pues la distribución de interés π sólo es usada a través del cociente $\pi(\theta')/\pi(\theta^{(k)})$ en (1.15).

Utilizando este esquema de muestreo es posible determinar diferentes algoritmos de actualización, dependiendo de la definición de la distribución q por utilizar. El algoritmo original considera un esquema de muestreo independiente, i.e. $q(\theta^{(k)},\theta') = q(\theta')$, en cuyo caso el cociente en (1.15) se reduce al cociente $w(\theta')/w(\theta^{(k)})$, donde $w(\theta) = \pi(\theta)/q(\theta)$ denota los pesos de importancia definidos previamente para aproximar integrales empleando como distribución de importancia a q. Otra alternativa consiste en definir q como una distribución simétrica, i.e. $q(\theta^{(k)},\theta') = q(\theta',\theta^{(k)})$, en cuyo caso el cociente en (1.15) se simplifica de la forma $\pi(\theta')/\pi(\theta^{(k)})$. Otra posibilidad consiste en definir $q(\theta^{(k)},\theta')$ a través de la relación $\Theta = \theta^{(k)} + Z$, donde Z es una variable aleatoria con media cero y función de distribución r. En este caso q se define a través de una caminata aleatoria, de manera que el algoritmo se concentra en explorar vecindades contiguas al estado previo de la cadena en el espacio de estados de la cadena. Es deseable que la distribución r sea simétrica. Para el caso de espacios no acotados, la distribución Normal (multivariada) o t son dos alternativas útiles y simples.

1.4.5. Muestreador de Gibbs

En algunas ocasiones tenemos el interés de obtener una muestra de una distribución $\pi(\cdot)$ multivariada. De esta forma, para obtener una muestra de $\pi(\cdot)$ mediante MCCM es necesario construir una cadena de Markov con un espacio de estado multivariado. En este caso el muestreador de Gibbs resulta un método práctico para construir tales cadenas preservando las características antes mencionadas.

Para estos efectos supongamos que la distribución de interés π corresponde a una variable aleatoria p-dimensional Θ, y que por razones prácticas podemos descomponer este espacio en $q \leq p$ componentes, denotados por $\Theta_1,...,\Theta_q$, algunos de éstos posiblemente multivariados, y denotemos por Θ_{-l} a los componentes de Θ menos el l-ésimo, para $l = 1,2,...,q$.

Dados los valores de la cadena en la iteración k, $\theta^{(k)} = (\theta_1^{(k)},...,\theta_q^{(k)})$, producimos la transición al estado $\theta^{(k+1)}$ mediante un muestreo sucesivo de las distribuciones condicionales completas mediante el siguiente

esquema:

$$\begin{aligned} \theta_1^{(k+1)} &\sim \pi(\theta_1|\theta_2^{(k)},...,\theta_q^{(k)}) \\ \theta_2^{(k+1)} &\sim \pi(\theta_2|\theta_1^{(k+1)},\theta_3^{(k)},...,\theta_q^{(k)}) \\ &\vdots \\ \theta_q^{(k+1)} &\sim \pi(\theta_q|\theta_1^{(k+1)},...,\theta_{q-1}^{(k+1)}). \end{aligned} \tag{1.16}$$

La estructura de actualización de los componentes dentro del algoritmo (1.16) puede definirse de manera aleatoria o determinista, considerando que cada componente es actualizado en cada ciclo al menos una vez. De esta forma la transición del estado $\theta^{(k)}$ al estado $\theta^{(k+1)}$ está determinada por:

$$K(\theta^{(k)},\theta^{(k+1)}) = \prod_{l=1}^{q} \pi(\theta_l^{(k+1)}|\theta_j^{(k+1)}, j<l, \theta_l^{(k)}, j>l).$$

En este algoritmo debe de considerarse la estructura de dependencia de los componentes individuales, y es recomendable agrupar en un bloque a aquellos componentes escalares que estén altamente correlacionados, para evitar que la cadena retarde su entrada al periodo de estabilidad. El muestreador de Gibbs puede ser visto como un caso particular del algoritmo de Metropolis-Hastings. En este caso el proceso de actualización se realiza en cada uno de los q componentes de Θ de la manera antes mencionada, entonces el valor propuesto de la cadena para el siguiente estado en cada l-ésimo componente es actualizado, con probabilidad 1, de la densidad $\pi(\theta_l^{(k+1)}|\theta_j^{(k+1)}, j<l, \theta_l^{(k)}, j>l)$.

Para implementar el muestreador de Gibbs de manera directa es necesario conocer de manera cerrada cada uno de las distribuciones condicionales completas y tener la capacidad de muestrear datos de ellas también directamente. En algunas ocasiones no es posible obtener de manera cerrada algunas de las distribuciones condicionales completas, que se conocen salvo su constante de normalización. En este caso, es posible diseñar el muestreador de Gibbs incorporando en la etapa de muestreo de la condicional no normalizada la generación de muestras de una distribución instrumental e incorporando ésta al proceso de muestreo mediante un paso adicional de importancia para la aceptación de esta muestra. Este algoritmo se conoce como muestreador de Gibbs por Importancia. Supongamos, sin pérdida de generalidad, que no es posible obtener la distribución condicional completa del j-ésimo componente en (1.16), y que podemos definir una distribución instrumental $q(\theta_j|\theta_{-j})$ completamente determinada que aproxima a $\pi(\theta_j|\theta_{-j})$. Esta distribución se deriva de la distribución instrumental $q(\theta)$ que aproxima a la distribución final completa $\pi(\theta)$. Así, en la j-ésima etapa de muestreo correspondiente de (1.16), donde el estado parcialmente actualizado

de la cadena es $\theta_{-j} = (\theta_1^{(k+1)}, ..., \theta_{j-1}^{(k+1)}, \theta_{j+1}^{(k)}, ..., \theta_q^{(k)})'$, para la $k+1$-ésima iteración de la cadena, definimos el cociente de importancia $w(\theta) = \pi(\theta)/q(\theta)$ y aceptamos la muestra $\theta_j^{'}$, generada por $q(\theta_j|\theta_{-j})$, con una probabilidad $\alpha(\theta, \theta^{'}) = \text{mín}\{1, w(\theta^{'})/w(\theta)\}$, donde $\theta^{'} = (\theta_1^{k+1}, ..., \theta_{j-1}^{(k+1)}, \theta_j^{'}, \theta_{j+1}^{(k)}, ..., \theta_q^{(k)})'$ y $\theta = (\theta_1^{k+1}, ..., \theta_{j-1}^{(k+1)}, \theta_j^{(k)}, \theta_{j+1}^{(k)}, ..., \theta_q^{(k)})'$, en otro caso $\theta_j^{(k+1)} = \theta_j^{(k)}$. Los pesos $w(\theta)$ en la probabilidad de aceptación son los mismos que empleamos para aproximar integrales mediante el muestreo por importancia, previamente discutido, así que las consideraciones presentadas para la distribución instrumental q tienen el mismo significado en este esquema. El componente aleatorio de aceptación garantiza que la cadena tenga a π como distribución invariante.

Los esquemas de muestreo que describimos previamente representan dos alternativas flexibles para implementar procedimientos Bayesianos de inferencia. Para problemas donde no puedan ser empleados de manera directa, se pueden definir diferentes combinaciones de éstos o de algunas generalizaciones, en diferentes etapas y bajo ciertas restricciones. Esta combinaciones dan origen a lo que se conoce como métodos híbrido de MCCM (**?**).

Por otro lado, éstos esquemas de muestreo están diseñados para generar muestras o una cadena de Markov, de una distribución π definida sobre un espacio de dimensión fija. En la siguiente sección describimos un método de muestreo diseñado medidas de probabilidad π definidas sobre un espacio de dimensiones cambiantes, en cuyo caso se generan muestras de medidas de probabilidad degeneradas en subespacios del espacio general de interés.

1.4.5.1. Consideraciones Generales

Existen diferentes consideraciones que se deben tomar en cuenta al momento de implementar algún algoritmo de muestreo mediante MCCM. Supongamos que $\{\theta^{(k)}\}_{k\geq 1}$ es una trayectoria de una cadena de Markov con distribución invariante π. La primera consideración sobre el uso de MCCM es determinar cuándo la cadena de la cual estamos simulando entra en su fase de equilibrio. Esta no es una tarea simple, y empíricamente es difícil asegurar este comportamiento, sin embargo existen diferentes método para su monitoreo, como por ejemplo, graficar la trayectoria o traza de la cadena y los promedios actualizados de cada uno de sus componentes. En la práctica es usual definir un periodo inicial de longitud considerable de manera que la cadena presumiblemente entre en su fase de equilibrio, para tratar de garantizar que la cadena no se afecte por el valor inicial de la cadena. Por otro lado, suponiendo que la cadena se encuentra dentro de su fase de

equilibrio, es evidente que debido a la estructura de Markov los datos muestreados no son independientes. Para reducir este efecto podemos obtener submuestras espaciadas de la trayectoria de la cadena simulada, de manera que la autocorrelación entre los datos no sea significativa. La longitud del submuestreo es arbitraria y se determina a partir de un análisis exploratorio de la trayectoria de la cadena. Alternativamente, para garantizar una muestra independiente de π, es posible generar un gran número de cadenas de manera simultánea y conservar los valores observados de cada cadena después de un periodo o longitud adecuada, de manera que la cadena este en su fase de equilibrio. Esta alternativa es poco eficiente ya que implica un costo computacional demasiado elevado durante su implementación.

Referencias