Clase 5: Inferencia bayesiana (parte 1)

MDS7104 Aprendizaje de Máquinas

Felipe Tobar

Iniciativa de Datos e Inteligencia Artificial
Universidad de Chile

7 de abril de 2024

UNIVERSIDAD
DE CHILE

Limitaciones del enfoque de máxima verosimilitud

- Una limitación del paradigma de MV es que no da la posibilidad de incorporar conocimiento experto, es decir, introducir sesgos sobre el parámetro al realizar inferencia.

- El paradigma bayesiano busca solucionar este problema interpretando el parámetro θ como variable aleatoria, donde la disponibilidad de datos se interpreta como un evento que aporta evidencia sobre el valor de θ.

- Consecuentemente, el proceso de inferencia ahora se centra en encontrar la distribución condicional:

$$p(\theta|\mathcal{D}),$$

y no simplemente $\arg\max_\theta p(\mathcal{D}|\theta)$.

Teorema de Bayes en el aprendizaje automático

Del curso de probabilidades, sabemos que el teorema de Bayes permite intercambiar la variable aleatoria y la cláusula condicionante de una distribución condicional. Asumiendo que el parámetro ahora es una variable aleatoria $\theta \sim p(\theta)$, de acuerdo al T. de Bayes se tiene que:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta),$$

donde $x \in \mathcal{X}$ es una observación, y $\theta \in \Theta$ es un parámetro. En la expresión anterior podemos identificar las siguientes cantidades:

- el prior o densidad a priori: $p(\theta)$.
- la verosimilitud: $p(x|\theta)$.
- la densidad posterior: $p(\theta|x)$.
- la densidad marginal de x: $p(x) = \int_\Theta p(x|\theta)p(\theta)\mathrm{d}\theta$.

En el problema de inferencia, los datos (x) son conocidos y fijos, mientras que el parámetro es desconocido (y por ende, aleatorio). Por esta razón, podemos considerar una versión proporcional de la posterior, omitiendo la distribución de los datos, pues solo importa la posterior como función de θ.

Nota: Incertidumbre epistemológica y aleatoria

Notemos que, a diferencia de MV, aquí θ es aleatorio por lo que (Θ, Σ, p) debe ser un espacio de probabilidad.

Elementos de la inferencia bayesiana

De la forma proporcional del teorema de Bayes, se reconocen dos elementos que constituyen el enfoque bayesiano:

- **Verosimilitud $p(x|\theta)$:** que modela la aleatoriedad del modelo, el cual produce datos de forma aleatoria incluso cuando el modelo es perfectamente conocido. Este tipo de incertidumbre no puede ser reducida observando datos.

- **La distribución a priori $p(\theta)$:** encapsula la incertidumbre epistemológica (lo que no sabemos) sobre el sistema, la cual pude ser reducida observando datos y calculando $p(\theta|x)$ mediante el Teorema de Bayes. Distintas distribuciones a priori llevarán a distintas distribuciones a posteriori, lo cual establece la subjetividad del enfoque bayesiano.

Nota: $p(\theta, x) = p(x|\theta)p(\theta)$

Uno de los problemas que tuvo este enfoque al comienzo viene dado por la subjetividad introducida en el prior $p(\theta)$. Si bien se podría proponer un prior uniforme $p(\theta) \propto 1$, este introduce sesgos de todas formas ya que no es invariante bajo transformaciones inyectivas. Actualmente, existen priors no informativos invariantes bajo inyecciones como por ejemplo el prior de Jeffreys.

Elección del prior: conjugación

El prior sobre el parámetro $p(\theta)$ encapsula toda la información experta que se le quiera dar al modelo. De esta forma, de acuerdo a $p(\theta|x) \propto p(x|\theta)p(\theta)$, el prior funciona como un ponderador de la verosimilitud de acuerdo a la importancia que se le dé a cada θ.

Con el modelo $p(x|\theta)$ acordado, solo queda elegir la distribución a priori, los cual es guiado por dos objetivos:

- En primer lugar, debemos encapsular lo que efectivamente que sabemos del parámetro θ.
- El segundo objetivo es obtener una forma *amigable* de la distribución posterior, en el sentido que esta sea un distribución con propiedades que deseemos, en particular, que la podamos calcular, evaluar, y samplear de ella.

Elección del prior: conjugación

El uso de un prior arbitrario resulta en que la posterior tenga una forma arbitraria también, con lo que, incluso si tanto el prior como la verosimilitud tienen formas *conocidas*, no tenemos ninguna garantía de que el posterior también la tenga y consecuentemente sea difícil de interpretar y calcular.

Una práctica usual es elegir un prior $p(\theta)$ tal que esté en la misma familia que la distribución posterior $p(\theta|\mathcal{D})$:

Definition (prior conjugado)

Diremos que el prior $p(\theta)$ es *conjugado* a la verosimilitud $p(D|\theta)$, cuando la posterior $p(\theta|\mathcal{D})$ está en la misma familia, es decir, tienen la misma distribución con parámetros distintos.

Una ventaja de utilizar priors conjugados es que la actualización de prior a posterior al actualizar la verosimilitud debido a la incorporación de datos, es simplemente un cambio de parámetros, lo que permite:

- poder interpretar los nuevos parámetros de acuerdo al modelo inicial
- que la actualización Bayesian sea equivalente a moverse en un espacio de dimensión fija (e.g. $\mathbb{R}^p$)

Prior conjugado sobre el modelo gaussiano

Consideremos un conjunto de observaciones $\mathcal{D} = \{x_i\}_{i=1}^n \subset \mathbb{R}$ generadas independiente e idénticamente distribuidas (iid) por el modelo $\mathcal{N}(\mu, \sigma^2)$. Recordemos que la verosimilitud de la media y varianza respectivamente está dada por

$$L_{\mathcal{D}}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right). \tag{0.1}$$

A continuación veremos priors conjugados para esta verosimilitud para dos casos:

- la media μ es desconocida y la varianza σ^2 es conocida.
- la varianza σ^2 es desconocida y la media μ es conocida.
- El caso donde tanto la media como la varianza son desconocidas, queda propuesto.

Prior conjugado sobre el modelo gaussiano: μ es desconocido

Consideremos el prior sobre la media $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$ donde μ_0 y σ_0^2 son parámetros fijos (hiperparámetros) y por lo tanto conocidos. Bajo este prior y denotando los datos $\mathcal{D} = (x_1, \ldots, x_n)$, la posterior está dada por:

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\
&\propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \right).
\end{aligned}
$$

Reordenando los términos dentro de la exponencial, se tiene que:

$$p(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma_n^2}(\mu - \mu_n)^2 \right) \text{ donde } \sigma_n^2 \text{ y } \mu_n \text{ serán definidos en breve.}$$

como $p(\mu|\mathcal{D})$ debe integrar uno, la única densidad de probabilidad proporcional al lado derecho de la ecuación anterior es la Gaussiana de media μ_n y varianza σ_n^2.

Consecuentemente, se prueba que el prior elegido era efectivamente conjugado con la verosimilitud gaussiana ya que la posterior también tiene distribución gaussiana.

Prior conjugado sobre el modelo gaussiano: μ es desconocido

Se probó que la posterior está dada por la siguiente gaussiana:

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu; \mu_n, \sigma_n^2) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu - \mu_n)^2\right),$$

donde, luego de reordenar términos, la media y la varianza están dadas por

$$\mu_n = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}\right), \quad \text{donde } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Se observa que los parámetros de la posterior son combinaciones entre los parámetros del prior y los datos. Además, a medida que se tienen más datos, el prior va perdiendo importancia y comienzan a predominar las observaciones.

Prior conjugado sobre el modelo gaussiano: σ^2 es desconocido

Ahora procedemos con el siguiente prior para la varianza, llamado gamma-inverso:

$$p(\sigma^2) = \text{inv-}\Gamma(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2).$$

Con este prior, la posterior de la varianza toma la forma:

$$\begin{aligned} p(\sigma^2|\mathcal{D}) &\propto \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \right) \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{N/2+\alpha+1}} \exp\left(-\frac{1}{\sigma^2} \left(\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 + \beta \right) \right), \end{aligned}$$

donde nuevamente la proporcionalidad ha sido mantenida debido a la remoción de las constantes. Esta última expresión es proporcional a una distribución gamma-inversa con hiperparámetros α_n y β_n, es decir:

$$p(\sigma^2|\mathcal{D}) \sim \text{inv-}\Gamma(\sigma^2; \alpha_n, \beta_n) \text{ donde } \alpha_n = \frac{n}{2} + \alpha \text{ y } \beta_n = \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 + \beta.$$