EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Inferencia Aproximada.
Objetivo: Que veremos.
Lectura recomendada: Una explicación de la aproximación Laplace la puedes encontrar en 2.4.4 de [3] y 13.3 de [2].

1. INTRODUCCIÓN

En inferencia Bayesiana definimos un modelo conjunto para observaciones y las configuraciones del proceso generador de datos. Esto nos permite utilizar el teorema de Bayes para actualizar nuestro conocimiento sobre los parámetros que no conocemos por medio de

$$\pi(\theta|y) \propto \pi(y|\theta)\,\pi(\theta)\,. \tag{1}$$

El procedimiento de inferencia dentro de este marco es sencillo y prácticamente directo pues se traduce en reportar la distribución posterior de las configuraciones en luz de las observaciones que tengamos.

A lo largo de este curso hemos establecido que de alguna u otra forma lo que necesitamos es reportar valores esperados (que se traduce en poder resolver integrales) utilizando dicho estado de conocimiento actualizado.

En esta sección estudiaremos mecanismos para utilizar aproximaciones al proceso de inferencia basado en el lado derecho de eq. (1).

2. MUESTREO Y APROXIMACIONES

Hasta ahora lo que hemos visto son métodos de `aproximación de integrales`. En particular utilizando el `método Monte Carlo`. Hemos discutido que este es un método de estimación insesgado del cual se pueden esperar algunas propiedades bondadosas en el largo plazo.

Con los métodos de `simulación Markoviana` esperamos poder eliminar los problemas de complejidad computacional que usualmente se encuentran en aplicaciones. Por ejemplo, la incapacidad de utilizar generadores de números aleatorios para cualquiera que sea la distribución dada.

Los métodos Markovianos generan muestras, que esperamos sean, ligeramente correlacionadas y cuya distribución corresponda a la distribución que nos interesa.

2.1. Aproximación por curvatura

Existen alternativas para poder aproximar el problema de inferencia Bayesiana. Esto es particularmente importante cuando no podemos esperar a que nuestros muestreadores converjan.

En clase hemos discutido que con un conjunto suficientemente grande de datos la distribución posterior se *parece* a una Gaussiana. Parece natural poder, entonces, construir una aproximación con estas características. Es decir,

$$\pi(\theta|y) \approx \mathsf{Normal}\left(\theta|\hat{\theta}, \Sigma_{\hat{\theta}}\right)\,, \tag{2}$$

donde

$$\hat{\theta} = \mathsf{moda}(\theta|y), \qquad \Sigma_{\hat{\theta}} = \left[-\nabla^2_{\theta}\log\pi(\hat{\theta}|y)\right]^{-1}\,. \tag{3}$$

La aproximación utiliza información de primero y segundo orden de la distribución posterior. Es decir

$$\hat{\theta} = \arg\max_{\theta} \pi(\theta|) \,, \tag{4}$$

y Σ_θ nos da la información de la curvatura. Esta `aproximación cuadrática` se denomina `aproximación de Laplace`.

> Para que la aproximación de Laplace tenga sentido para un problema con parámetros restringidos, es usual transformar los parámetros a una escala sin restricciones. Por ejemplo, utilizando una transformación logarítmica o *logit* y recordando incorporar el término multiplicativo de la Jacobiana de dicha transformación.

La aproximación de Laplace nos permite sustituir un modelo de probabilidad por otro. Aunque en teoría podemos determinar la calidad de la aproximación –por medio de la expansión de Taylor– en la práctica puede resultar infactible dar una estimación de este error.

El problema de la aproximación de Laplace es que utiliza información local y puede fallar en capturar propiedades globales importantes de la distribución objetivo.

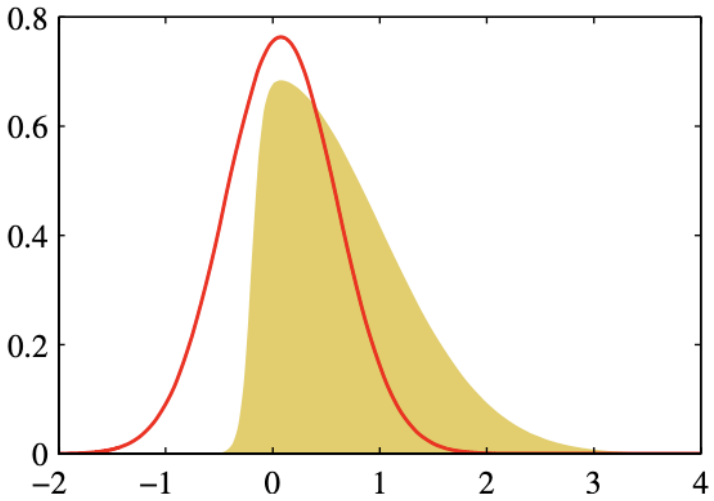

FIGURA 1. *Imagen tomada de [1]. Aproximación de Laplace.*

2.2. Aproximación por optimización

Una alternativa es definir una familia de posibles distribuciones $\mathcal{Q}$ y encontrar dentro de esta familia de distribuciones la que `mejor se parezca` a nuestra distribución objetivo $\pi(\theta|y)$.

Lo importante es poder definir la noción de encontrar al mejor candidato dentro de Q.

2.3. La solución

En inferencia aproximada buscamos sustituir nuestra distribución objetivo con aquella que resuelva el problema

$$\min_{q \in \mathcal{Q}} \mathsf{KL}\Big(q(\theta) \Big\| \pi(\theta|y) \Big) \,, \tag{5}$$

donde la familia de distribuciones $\mathcal{Q}$ define la calidad/complejidad de nuestra aproximación.

El problema es que no podemos calcular la divergencia de KL, pues necesitamos calcular la constante de normalización en $\pi(\theta|y)$. Así que lo que hacemos es re-expresar

$$\mathsf{KL}(q(\theta)\|\pi(\theta|y)) = \mathbb{E}[\log q(\theta)] - \mathbb{E}[\log \pi(\theta, y)] + \log \pi(y)\,. \tag{6}$$

El lado derecho en el primer término define

$$\mathsf{ELBO}(q) := \mathbb{E}[\log \pi(\theta, y)] - \mathbb{E}[\log q(\theta)]\,. \tag{7}$$

la cual se denomina la cota inferior de evidencia.

La cual podemos usar para re-expresar

$$\log \pi(y) = \mathsf{KL}(q(\theta)\|\pi(\theta|y)) + \mathsf{ELBO}(q)\,, \tag{8}$$

de donde podemos ver que lo que podemos buscar es `maximizar` el `ELBO` en lugar de `minimizar` la divergencia `KL`.

Nota que también podemos expresar

$$\mathsf{ELBO}(q) = \mathbb{E}[\log \pi(y|\theta)] - \mathsf{KL}(q(\theta)\|\pi(\theta))\,. \tag{9}$$

Lo cual nos dice que la distribución $q \in \mathcal{Q}$ que encontraremos será aquella que busque configuraciones afines al proceso generador de datos y que sea cercana a la distribución inicial.

En fig. 2 se muestra la solución encontrada minimizando el criterio de `ELBO`.

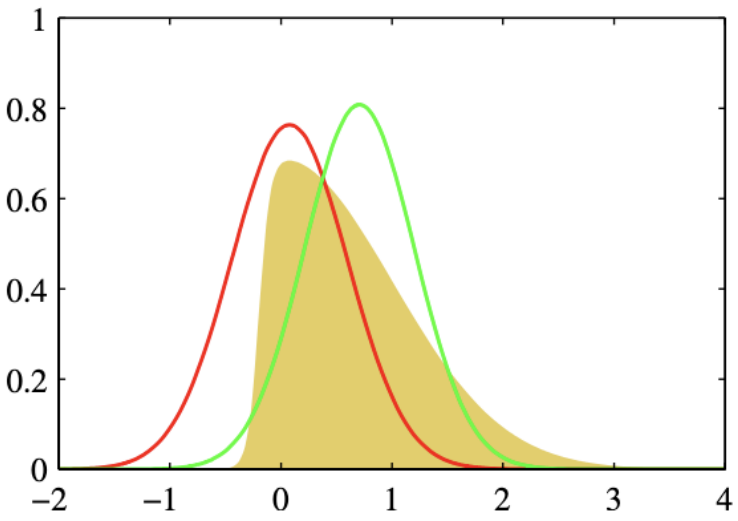

FIGURA 2. *Imagen tomada de [1]. Solución de `ELBO` se muestra en verde. Aproximación de Laplace en rojo.*

2.4. Dirección de `KL`

Hemos tomado la solución de $\mathsf{KL}(q\|\pi)$ por cuestiones numéricas y también discutimos que la solución tiene la interpretación de ser una aproximación de la posterior (justo lo que nos interesa).

Por ejemplo, en fig. 3, bajo una familia de Gaussianas independientes para $\mathcal{Q}$ la solución de $\mathsf{KL}(q\|\pi)$, además, se puede ver como una distribución que se concentra en las zonas de alta probabilidad. Mientras que la solución de $\mathsf{KL}(\pi\|q)$ se concentra en zonas de alta densidad. Lo que nos habla que la formulación correcta se fijará en las propiedades que nos interesen.

El mismo efecto se muestra en

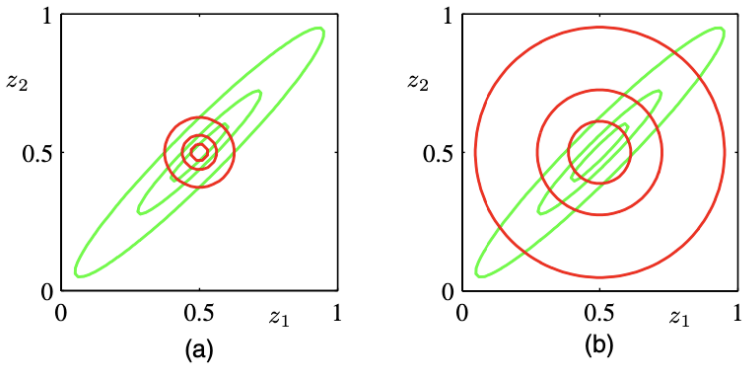

FIGURA 3. *Imagen tomada de [1]. En (a) se muestra* $\mathsf{KL}(q\|\pi)$ *y en (b) se muestra* $\mathsf{KL}(\pi\|q)$ *donde* $q \in \mathcal{Q}$ *y* π *es la distribución objetivo.*

2.5. Solución en la práctica

REFERENCIAS

[1] C. M. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2. 2, 3, 4

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL, 2014. 1

[3] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC Texts in Statistical Science. Taylor and Francis, CRC Press, Boca Raton, Second edition, 2020. ISBN 978-0-367-13991-9. 1