Clase 6: Inferencia bayesiana (parte 2)

MA5204 Aprendizaje de Máquinas

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

11 de abril de 2021

UNIVERSIDAD
DE CHILE

Recuerdo

Se vio que el enfoque bayesiano estaba basado en la relación dada por el teorema de Bayes

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

donde $p(\theta)$ era el prior sobre el parámetro θ y contenía el conocimiento experto.

Además, se estableció que un prior $p(\theta)$ es *conjugado* a la verosimilitud $p(D|\theta)$, cuando la posterior $p(\theta|D)$ está en la misma familia que el prior $p(\theta)$.

Se estudió el caso del modelo gaussiano, donde se llegó a la siguiente conclusión:

- μ **desconocido y** σ^2 **conocido:** el prior gaussiano para μ es conjugado con la verosimilitud gaussiana.
- σ^2 **desconocido y** μ **conocido:** el prior gamma-inverso para σ^2 es conjugado con la verosimilitud gaussiana.

Prior conjugado para el modelo lineal gaussiano

(regresón bayesiana)

Para un conjunto de observaciones $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, el modelo de regresión lineal puede ser escrito en forma vectorial utilizando la matriz de diseño trabajada en el capítulo de regresión:

$$Y = \tilde{X}\theta + \boldsymbol{\epsilon}.$$

De esta forma, la verosimilitud está dada por

$$\begin{aligned} L(\theta, \sigma^2) &= \text{MVN}(Y; \tilde{X}\theta, \mathbb{I}\sigma^2) \\ &\propto \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(Y - \tilde{X}\theta)^\top (Y - \tilde{X}\theta)\right). \end{aligned}$$

Esta última expresión es proporcional a una distribución gamma-inversa para σ^2 y proporcional a una MVN para θ. Consecuentemente, esta verosmilitud tiene los mismos priors conjugados que el modelo gaussiano vistos en la clase anterior.

Prior conjugado para el modelo lineal gaussiano

(regresón bayesiana)

Consideremos el caso en que σ^2 es conocido, por lo que elegimos el prior gaussiano para θ dado por

$$p(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(\theta-\theta_0)^\top \Lambda_0 (\theta-\theta_0)\right)$$

Entonces, de forma análoga al caso gaussiano, se obtiene una distribución posterior **para el parámetro θ de la regresión lineal** dada por $\text{MVN}(\theta; \theta_n, \sigma^2\Lambda_n^{-1})$, donde

$$\theta_n = (\tilde{X}^\top\tilde{X} + \Lambda_0)^{-1}(\tilde{X}^\top Y + \Lambda_0\theta_0) = (\tilde{X}^\top\tilde{X} + \Lambda_0)^{-1}(\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top Y + \Lambda_0\theta_0)$$

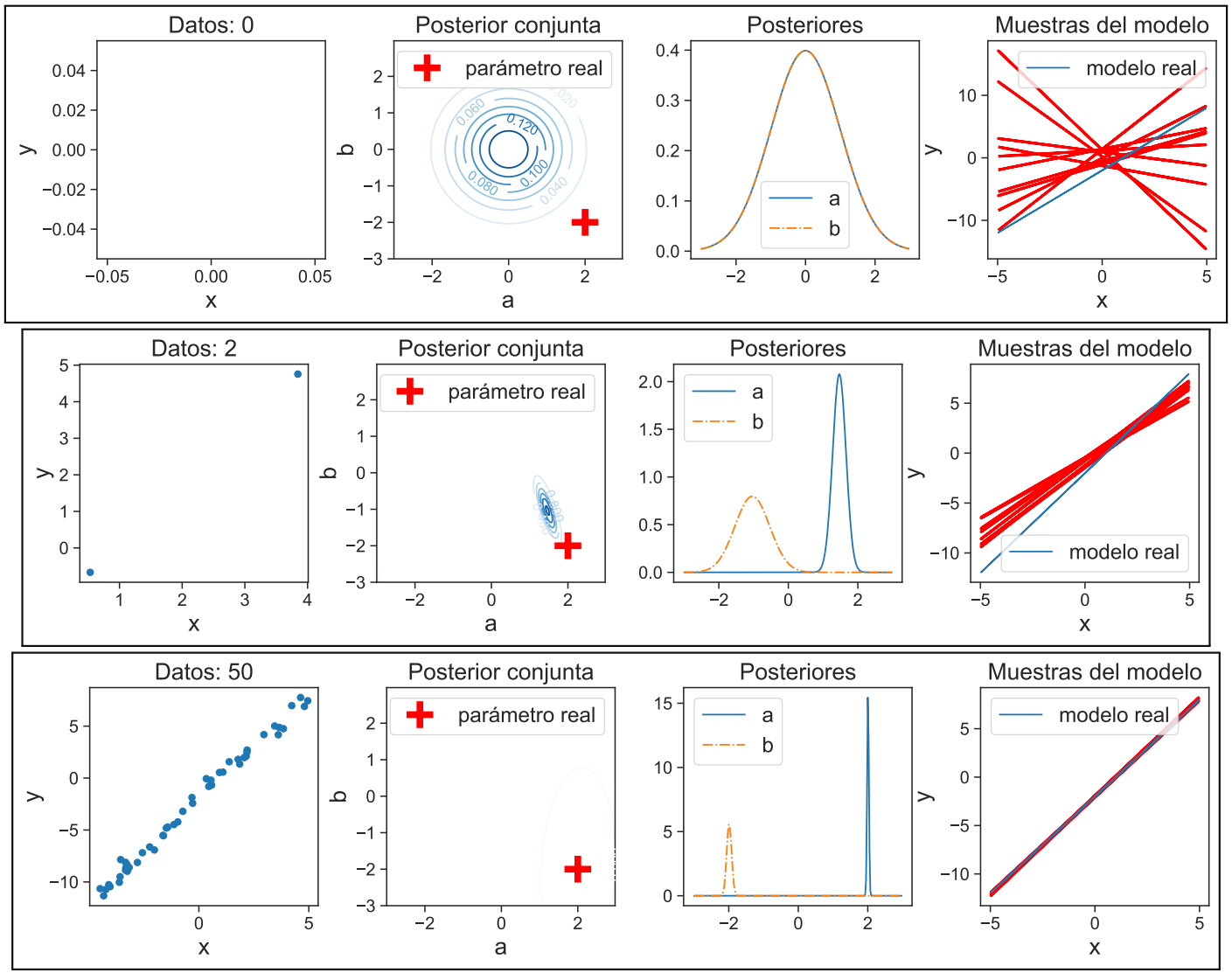
$$\Lambda_n = (\tilde{X}^\top\tilde{X} + \Lambda_0)$$

Es decir:

- La media posterior θ_n es un promedio ponderado entre la media a priori θ_0 y el estimador de máxima verosimilitud $(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top Y$.
- La varianza $\Sigma_n = \Lambda_n^{-1}$ se mueve desde la varianza a priori Λ_0^{-1} hacia $(\tilde{X}^\top\tilde{X})^{-1}$ a medida recibimos más observaciones, resultando en un modelo más preciso.

Regresión linea bayesiana: ejemplo

Implementación de la regresión lineal bayesiana, para $y = 2x - 2 + \epsilon$, con $\epsilon \sim \mathcal{N}(0, 0.5^2)$.



Máximo a posteriori (MAP)

Existen distintas formas de extraer una estimación puntual de un parámetro θ a partir de la distribución posterior $p(\theta|\mathcal{D})$ como por ejemplo, la moda, media o mediana, los cuales son equivalentes cuando la posterior es gaussiana.

Siguiendo un criterio similar al de máxima verosimilitud consideraremos estimaciones puntuales mediante la maximización de la distribución posterior, resumiendo la información de la posterior mediante su moda.

Definition (Máximo a posteriori)

Sea $\theta \in \Theta$ un parámetro con distribución posterior $p(\theta|\mathcal{D})$ definida en todo Θ, entonces nos referimos a estimación puntal dada por

$$\theta_{\text{MAP}} = \arg\max_{\Theta} p(\theta|\mathcal{D}),$$

como *máximo a posteriori (MAP)*.

La relación entre MAP y MV puede ser vista como un término adicional al momento de maximizar:

$$\theta_{\text{MAP}} = \arg\max_{\theta\in\Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta\in\Theta} p(\mathcal{D}|\theta)p(\theta) = \arg\max_{\theta\in\Theta} \left(\underbrace{\log p(\mathcal{D}|\theta)}_{l(\theta)} + \log p(\theta) \right)$$

MAP para el modelo lineal gaussiano

Para el modelo lineal y gaussiano, se puede considerar un prior gaussiano multivariado donde cada coordenada de θ tendrá un prior independiente de media cero y varianza σ_θ^2.
Asumiendo la varianza del ruido σ_ϵ^2 conocida, se puede calcular θ_{MAP} como:

$$
\begin{aligned}
\theta^\star_{\text{MAP}} &= \arg\max_\theta p(Y|\tilde{X},\theta)p(\theta) \\
&= \arg\max_\theta \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left[\frac{-1}{2\sigma_\epsilon^2}(y_i - \theta^\top \tilde{x}_i)^2 \right] \right) \frac{1}{(\sqrt{2\pi}\sigma_\theta)^{M+1}} \exp\left(\frac{-\theta^\top \theta}{2\sigma_\theta^2} \right) \\
&= \arg\max_\theta \frac{1}{(\sqrt{2\pi}\sigma_\epsilon)^N} \frac{1}{(\sqrt{2\pi}\sigma_\theta)^{M+1}} \exp\left(\sum_{i=1}^{N} \frac{-1}{2\sigma_\epsilon^2}(y_i - \theta^\top \tilde{x}_i)^2 - \frac{||\theta||^2}{2\sigma_\theta^2} \right) \\
&= \arg\min_\theta \sum_{i=1}^{N} (y_i - \theta^\top \tilde{x}_i)^2 + \frac{\sigma_\epsilon^2}{\sigma_\theta^2}||\theta||^2
\end{aligned}
$$

Esta expresión es equivalente al funcional de ridge regression, es decir, la solución MAP del modelo lineal y gaussiano con prior gaussiano es equivalente a la de mínimos cuadrados regularizados con orden de regularización $p = 2$.

Consideraciones generales

- En el caso anterior, si se hubiese elegido un prior exponencial $p(\theta) \propto \exp(-\gamma|\theta|)$, se hubiese llegado llegado a MCR con regularización $p = 1$ (LASSO).
- Desde ahora, nos referirnos como MAP a **cualquier** estimación puntual, pues, como acabamos de ver, esta es equivalente a MCR y al mismo tiempo equivale al criterio de máxima verosimilitud cuando el prior es uniforme.
- Para modelos generales (distintos al caso lineal y gaussiano) el MAP no podrá ser calculado de forma explícita imponiendo

$$\nabla_\theta \log p(\theta|\mathcal{D}) = 0,$$

sino que se tendrán que considerar algoritmos de optimización. En particular se utilizan algoritmos basados en derivadas con iteraciones de la forma

$$\theta_{i+1} = \theta_i + \eta \nabla_\theta \log p(\theta_i|\mathcal{D}),$$

donde se espera que la secuencia $\{\theta_i\}_{i\in\mathbb{N}}$ converja, es decir, $\nabla_\theta \log p(\theta_i|\mathcal{D}) \to 0$.