Estadística Bayesiana
Parte IV

Irving Gómez Méndez

Evaluación y Comparación de Modelos

Sean p y q las distribuciones predictivas posteriores bajo dos modelos distintos $\mathcal{P}$ y $\mathcal{Q}$, y suponga que la verdadera densidad de nuestros datos es f. Vamos a preferir el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si la divergencia KL entre f y p es menor que la divergencia KL entre f y q.

Es decir, preferimos $\mathcal{P}$ sobre $\mathcal{Q}$ si

$$KL(f||p) < KL(f||q)$$

$$\begin{aligned} \Leftrightarrow & \mathbb{E}_{Y\sim f}[\log f(Y)] - \mathbb{E}_{Y\sim f}[\log p(Y|\mathbf{Y})] \\ & < \mathbb{E}_{Y\sim f}[\log f(Y)] - \mathbb{E}_{Y\sim f}[\log q(Y|\mathbf{Y})] \end{aligned}$$

$$\Leftrightarrow \mathbb{E}_{Y\sim f}[\log p(Y|\mathbf{Y})] > \mathbb{E}_{Y\sim f}[\log q(Y|\mathbf{Y})]$$

Sin embargo, todavía tenemos el problema de tener que calcular un valor esperado con respecto a la verdadera distribución f. Para solucionar este problema podemos hacer uso una vez más de los datos $Y_1, \ldots, Y_n$ que sabemos que tienen densidad f.

Por lo tanto, preferiremos el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i|\mathbf{Y}) > \frac{1}{n}\sum_{i=1}^{n}\log q(Y_i|\mathbf{Y}) \\ \Leftrightarrow &\sum_{i=1}^{n}\log p(Y_i|\mathbf{Y}) > \sum_{i=1}^{n}\log q(Y_i|\mathbf{Y}) \end{aligned}$$

Por otro lado, recuerde que la distribución predictiva posterior puede ser escrita de la siguiente manera

$$\begin{aligned} p(Y|\mathbf{Y}) &= \int_{\Theta} p(Y, \theta|\mathbf{Y}) d\theta \\ &= \int_{\Theta} p(Y|\theta) p(\theta|\mathbf{Y}) d\theta, \end{aligned}$$

por lo tanto, preferiremos el modelo asociado a p sobre el modelo asociado a q si

$$\sum_{i=1}^{n} \log \int_{\Theta} p(Y_i|\theta) p(\theta|\mathbf{Y}) d\theta > \sum_{i=1}^{n} \log \int_{\Theta} q(Y_i|\theta) q(\theta|\mathbf{Y}) d\theta$$

A

$$\sum_{i=1}^{n} \log \int_{\Theta} p(Y_i|\theta)p(\theta|\mathbf{Y})d\theta$$

se le llama *log pointwise predictive density* (lppd).

Si contamos con una muestra $\theta_1, \ldots, \theta_S$ de la distribución posterior de θ, entonces podemos aproximar $\int_{\Theta} p(Y|\theta)p(\theta|\mathbf{Y})d\theta$ de la siguiente manera

$$\int_{\Theta} p(Y|\theta)p(\theta|\mathbf{Y})d\theta \approx \frac{1}{S}\sum_{s=1}^{S} p(Y|\theta_s)$$

Por lo tanto, preferiremos el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si

$$\sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right] > \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} q(Y|\theta_s) \right].$$

A

$$\sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right]$$

la llamaremos *computed log pointwise predictive density* (computed lppd).

Note que si S es lo suficientemente grande entonces computed lppd aproximará muy bien a lppd. Por esta razón algunos autores definen el lppd como esta segunda expresión.

Aunque, en principio necesitamos calcular la densidad de cada observación Y_i en cada muestra del parámetro θ_s, en la práctica podemos enfrentar problemas numéricos, por lo que es preferible reescribir el *computed lppd* de la siguiente manera:

$$\begin{aligned}
\text{computed lppd} &= \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right] \\
&= \sum_{i=1}^{n} \left[\log \sum_{s=1}^{S} p(Y_i|\theta_s) - \log(S) \right] \\
&= \sum_{i=1}^{n} \left[\log \sum_{s=1}^{S} \exp \left\{ \log p(Y_i|\theta_s) \right\} - \log(S) \right]
\end{aligned}$$

La función $\log \sum \exp \{\cdot\}$ suele estar programada en varios lenguajes de programación de manera eficiente, manteniendo la precisión númerica, normalmente se le llama logsumexp. Así, podemos calcular el *computed lppd* como:

$$\text{computed lppd} = \sum_{i=1}^{n} [\text{logsumexp}(\log p(Y_i|\theta_s)) - \log(S)]$$

El caso ideal para evaluar un modelo sería contar con una muestra de entrenamiento, con la que se ajusta el modelo; y una muestra de prueba, con la que se calcula la métrica de desempeño del modelo.

Es decir, suponga que además de nuestra muestra de entrenamiento $Y_1, \ldots, Y_n$, también contamos con una muestra de prueba $\tilde{Y}_1, \ldots, \tilde{Y}_m$, con la que calcularíamos nuestra métrica

$$\text{computed lppd} = \sum_{i=1}^{m} \Big[\text{logsumexp}(\log p(\tilde{Y}_i|\theta_s)) - \log(S)\Big]$$

Sin embargo, cuando no contamos con una muestra de prueba, debemos tener en cuenta que estamos usando dos veces nuestros datos. Una primera vez para ajustar el modelo y obtener una muestra de la posterior, y una segunda vez para evaluar el modelo. Esto provocará una evaluación más optimista del modelo.

El plan es entonces calcular el lppd y después agregar algún término de penalización que corrija la estimación y evitar así elegir modelos sobreajustados.

Criterios de Información

Por razones históricas a las medidas de exactitud predictiva se les llama criterios de información, y típicamente se definen en términos de la devianza. Es importante aclarar que no existe un único concenso al definir los criterios de información y, por lo tanto, sus definiciones pueden variar ligeramente.

Devianza

Típicamente se define la devianza como -2 veces la logverosimilitud de los datos fijando los parámetros en algún valor, es decir $-2 \log p(\mathbf{Y}|\hat{\theta})$. Aunque desde un enfoque puramente bayesiano, algunos autores la definen como -2 veces el lppd. El -2 es simplemente por razones históricas.

Criterio de la información de Akaike (AIC)

El criterio de información más conocido es el criterio de la información de Akaike (AIC). La corrección más sencilla está basada en el comportamiento asintótico normal de la distribución posterior.

Sea k el número de parámetros estimados en el modelo. En el caso límite (o en casos especiales como cuando se cuenta con un modelo normal lineal con varianza conocida y previa uniforme), restar k de la logverosimilitud dado el estimador de máxima verosimilitud es la manera de corregir la sobreestimación del poder predictivo del modelo

$$\log p(\mathbf{Y}|\hat{\theta}_{mle}) - k.$$

El AIC se define como la expresión anterior multiplicada por -2,

$$\text{AIC} = -2 \log p(\mathbf{Y}|\hat{\theta}_{mle}) + 2k.$$

Desde el enfoque bayesiano, algunos autores redefinen el AIC como

$$\text{AIC} = -2\mathsf{lppd} + 2k.$$

Cuando ya no se cuenta con un modelo lineal con previas uniformes se vuelve inapropiado simplemente agregar el número de parámetros. Para modelos con previas informativas o con una estructura jerárquica, el número efectivo de parámetros depende de la varianza de los parámetros al nivel del grupo.

Criterio de información de la devianza (DIC)

El DIC es una especie de versión bayesiana del AIC, haciendo dos cambios. El primero es reemplazando el estimador de máxima verosimilitud por la media a posteriori $\hat{\theta}_{Bayes} = \mathbb{E}(\theta|\mathbf{Y})$ y el segundo es que reemplaza a k por el número efectivo de parámetros p_{DIC}.

El criterio de la información de la devianza se define entonces como

$$\text{DIC} = 2\log p(\mathbf{Y}|\hat{\theta}_{\text{Bayes}}) + 2p_{\text{DIC}}.$$

Existen al menos dos maneras distintas de definir el número efectivo de parámetros:

$$p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}(\log p(\mathbf{Y}|\theta))\right),$$

usando una muestra de la posterior $\theta_1, \ldots, \theta_S$, esta expresión puede ser aproximada por

$$\text{computed}p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S}(\log p(\mathbf{Y}|\theta_s))\right).$$

Si la media posterior se encuentra lejos del máximo a posteriori, existe el problema de que p_{DIC} tome un valor negativo.

Una versión alternativa para el número efectivo de parámetros está dada por:

$$p_{\text{DIC alt}} = 2\mathbb{V}_{\theta\sim p(\theta|\mathbf{Y})}\left(\log p(\mathbf{Y}|\theta)\right),$$

usando una muestra de la posterior $\theta_1, \ldots, \theta_S$, esta expresión puede ser aproximada por

$$\mathsf{computed}p_{\text{DIC alt}} = 2\mathbb{V}_{s=1}^{S} \log p(\mathbf{Y}|\theta_s),$$

donde $\mathbb{V}_{s=1}^{S}$ representa la varianza muestral

$$\mathbb{V}_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$$

Criterio de la información de Watanabe-Akaike (WAIC)

El criterio de la información de Watanabe-Akaike o *widely applicable information criterion* (WAIC) también acepta al menos dos maneras distintas de definir el número efectivo de parámetros

$$p_{\text{WAIC 1}} = 2\sum_{i=1}^{n}\left(\log(\mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}p(Y_i|\theta)) - \mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}(\log p(Y_i|\theta))\right).$$

$$\mathsf{computed}p_{\text{WAIC 1}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S}p(Y_i|\theta_s)\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(Y_i|\theta_s)\right).$$

La segunda manera de definir el número efectivo de parámetros está dada por

$$p_{\text{WAIC 2}} = \sum_{i=1}^{n} \mathbb{V}_{\theta \sim p(\theta|\mathbf{Y})} \log p(Y_i|\theta_s),$$

$$\text{computed} p_{\text{WAIC 2}} = \sum_{i=1}^{n} \mathbb{V}_{s=1}^{S} \log p(Y_i|\theta_s).$$

Así, se define el WAIC como

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}.$$

Validación cruzada

También se puede corregir el optimismo generado al usar dos veces los datos aplicando validación cruzada. Sin embargo, validación cruzada puede ser computacionalmente intensiva, pues requiere varias particiones de los datos. En un caso extremo *leave-one-out cross-validation* (LOO-CV) requiere n particiones de los datos.

PSIS

Sin embargo, existen soluciones ingeniosas que permiten aproximar validación cruzada sin tener que ajustar el modelo varias veces. Una de las soluciones es usando la "importancia" de cada observación en la distribución posterior. Dicha "importancia" significa un mayor impacto en la distribución posterior, si removemos una observación importante, la posterior cambiará más. Este método es llamado *Pareto-smoothed importance sampling cross-validation* (PSIS).