Introducción a la Probabilidad Bayesiana. Parte 6, Comparación de modelos, la navaja de Occam y los P-values.

Martín Onetto

Bayes Factor y la navaja de Occam

Hasta ahora nos hemos concentrado en actualizar nuestra información sobre los parámetros $\boldsymbol{\theta}$. Pero qué sucede cuando tenemos más de un modelo M_i para describir a nuestros datos y consideramos a cada uno de ellos como válidos. Para cada uno de ellos tenemos:

$$P(M_i|D,I) = \frac{P(D|M_iI)P(M_i|I)}{P(D|I)} \tag{1}$$

Si retomamos la interpretación bayesiana de la probabilidad, la expresión $P(M_i|D,I)$ cuantifica la confianza tenemos de que el modelo M_i sea cierto. Ahora bien, siguiendo la filosofía de la famosa expresión "*All models are wrong*" de George Box, el valor de $P(M_i|D,I)$ nunca será 1 en la práctica. Por lo tanto en lugar de tomar la probabilidad de un modelo de forma absoluta, podemos ver la relación entre las probabilidades de cada modelo. Para eso hacemos:

$$\frac{P(M_i|D,I)}{P(M_j|D,I)} = \frac{P(D|M_iI)}{P(D|M_jI)}\frac{P(M_i|I)}{P(M_j|I)} \tag{2}$$

al hacer el cociente la probabilidad de los datos sin especificar el modelo, $P(D|I)$, se cancela y nos queda simplemente el cociente entre la probabilidad de los datos D y entre las priors de los modelos.

El término $P(D|M_iI)$ corresponde a la constante de normalización de la posterior de los parámetros $\boldsymbol{\theta_i}$. Es decir:

$$P(D|M_i,I) = \int_{\theta_{iN}} ... \int_{\theta_{i1}} P(D|\boldsymbol{\theta_i},M_i,I)P(\boldsymbol{\theta_i}|M_i,I)\, d\theta_{i1}...d\theta_{iN} \tag{3}$$

Si hacemos la aproximación gaussiana sobre la likelihood, $\mathcal{L}(\theta) \equiv P(D|\boldsymbol{\theta_i},M_i,I)$ tenemos:

$$\log \mathcal{L}(\theta) \sim \log \mathcal{L}(\theta_{max}) + \frac{1}{2}\sum_{ij} (\boldsymbol{\theta}_{max} - \boldsymbol{\theta})_i \left(\frac{\partial^2 \mathcal{L}}{\partial\theta_i\theta_j}\right)_{\boldsymbol{\theta}_{max}} (\boldsymbol{\theta}_{max} - \boldsymbol{\theta})_j \tag{4}$$

Si reemplazamos ésto en la Eq. 3 obtenemos:

$$P(D|M_i, I) =$$
$$e^{\mathcal{L}(\theta_{max})} \int_{\theta_{iN}} ... \int_{\theta_{i1}} \exp\left\{ \frac{1}{2} \sum_{ij} (\boldsymbol{\theta}_{max} - \boldsymbol{\theta})_i \left(\frac{\partial^2 \mathcal{L}}{\partial \theta_i \theta_j} \right)_{\boldsymbol{\theta}_{max}} (\boldsymbol{\theta}_{max} - \boldsymbol{\theta})_j \right\} P(\boldsymbol{\theta_i}|M_i, I)\, d\theta_{i1} ... d\theta_{iN} \tag{5}$$

Acá vamos a tomar una simplificación para poder interpretar más el resultado. Consideraremos a los θ_i independiente entre sí, tal que la matriz $\left(\frac{\partial^2 \mathcal{L}}{\partial \theta_i \theta_j} \right)^{-1}_{\boldsymbol{\theta}_{max}}$ es diagonal con elementos $\delta\theta_i$, que representan el ancho de cada parámetro en la likelihood. Tomaremos una segunda simplificación diciendo que las priors de los parámetros son independientes entre sí, planas y acotadas, tal que

$$P(\boldsymbol{\theta}|M_i) = \prod_{\alpha} P(\theta_\alpha|M_i) = \prod_{\alpha} \frac{1}{\Delta\theta_\alpha} \tag{6}$$

donde cada $\frac{1}{\Delta\theta_\alpha}$ es la prior plana del parámetro θ_α. Como estas distribuciones priors son constantes, el valor $\Delta\theta_\alpha$ representa el ancho de la distribución. Reemplazando este resultado y haciendo la integral obtenemos:

$$P(D|M_i, I) = e^{\mathcal{L}(\theta_{max})} \prod_{\alpha} \frac{\delta\theta^i_\alpha}{\Delta\theta^i_\alpha} \tag{7}$$

Ahora si retomamos la ecuación 2 utilizando estos resultados:

$$\frac{P(M_i|D, I)}{P(M_j|D, I)} = \frac{P(D|M_i I)}{P(D|M_j I)} \frac{P(M_i|I)}{P(M_j|I)} = e^{\mathcal{L}(\theta^i_{max}) - \mathcal{L}(\theta^j_{max})} \frac{\prod_\alpha \delta\theta^i_\alpha / \Delta\theta^i_\alpha}{\prod_\beta \delta\theta^j_\beta / \Delta\theta^j_\beta} \frac{P(M_i|I)}{P(M_j|I)} \tag{8}$$

donde los índices α y β corresponde al número de parámetros que tiene el modelo i y j, respectivamente.

A los factores que no tienen que ver con la prior de los modelos $P(M|I)$ se los conoce como *Bayes factors* y se los suele notar como B_{ij} notando como subíndices que estamos comparando el modelo i con el j. Por lo tanto la comparación de modelos se resume en:

$$\frac{P(M_i|D, I)}{P(M_j|D, I)} = B_{ij} \frac{P(M_i|I)}{P(M_j|I)} \tag{9}$$

donde diferenciamos la información sobre los datos y los parámetros de la prior de los modelos.Ahora podemos observar más de cerca B_{ij},

$$B_{ij} = e^{\mathcal{L}(\theta^i_{max}) - \mathcal{L}(\theta^j_{max})} \times \frac{\Omega_i}{\Omega_j} \tag{10}$$

donde notamos $\Omega_k \equiv \prod_\gamma \delta\theta^k_\gamma/\Delta\theta^k_\gamma$ para $k = i, j$ y $\gamma = \alpha, \beta$.

El primer factor de la ecuación anterior compara simplemente el valor de las likelihoods en los máximos de cada modelo. Es decir, cuantifica cuánto más probable es haber medido los datos dado el modelo i comparado con el j.

Por otro lado tenemos el cociente entre los Ω de cada modelo, el cual representa la relación entre la incerteza que tenemos sobre un parámetro antes y después de haber considerado los datos. Recordemos que $\delta\theta$ corresponde al ancho de la likelihood y $\Delta\theta$ al ancho de la prior. En primera instancia esto nos dice que si introducimos un parámetro cuya información no cambia por los datos es invisible ante el procedimiento. Luego, si consideramos parámetros que sí se vuelven más certeros una vez que incorporamos la información de los datos, el Ω tendrá peso en la comparación entre M_i y M_j. Como observación general del factor Ω, vemos que hay un factor de penalización en introducir un parámetro del cual sabemos poco, es decir $\Delta\theta$ grande o prior ancha, y sólo será conveniente en el caso de que la probabilidad de los datos $e^{\mathcal{L}(\theta_{max})}$ se incremente aún más que el peso de este factor $\delta\theta/\Delta\theta$.

Visto de manera simple el hecho de que la comparación de modelos favorezca a modelos con menos parámetros está de acuerdo con cierta filosofía que se ha trasmitido en la comunidad científica a lo largo de los años, la idea de la *navaja de Occam* la cual favorece modelos que sean más simples en comparación a aquellos más complejos. Esta bella idea en general no exhibe una clara definición de qué es la simpleza, por lo que a pesar de querer ser tomada como principio metodológico peca de no poder ser consistente entre distintas personas dado que no está explícitamente definido. Sin embargo, bajo el procedimiento que mostramos a la hora de comparar modelos bajo la metodología bayesiana encontramos que el factor Ω_i/Ω_j nos hace una comparación cuantitativa sobre la "simplicidad" de cada modelo y la pesa según que tan bueno es el ajuste en caso de complejizarlo. Es por esto que se reconoce a Ω_i/Ω_j como el *factor de Occam*.

P-values tradicionales y Bayesianos

Una vez que logramos constuir un modelo probabilistico y calcular la distribución posterior de lo los parámetros que nos interesan, es importante evaluar cómo ajusta nuestro modelo a los datos y al conocimiento que tenemos del problema. En general es dificil incorporar toda la información que disponemos al modelado y es importante investigar qué aspectos **no** están siendo capturados

Como dijimos antes "All models are wrong", incluso el modelo de tirar de monedas o dados tiene problemas y sus resultados no son independientes. Por lo que preguntarnos si nuestro modelo es la verdad resulta una *mala* pregunta. Una pregunta más adecuada es "¿Las definciencias de nuestro modelo tienen efectos notables en las inferencias que produce?.º en layman terms "¿Las inferencias del modelo tienen sentido?"

En primer lugar nos vamos a concentrar en hacer esta evalución con los datos que ya tenemos, esto nos adelanta a cómo se va a comportar el modelo para futuros datos.

Si decimos que un modelo ajusta bien, significa que los dato generados con él deberían verse similar a los que observamos. En términos más técnicos los datos obervados deberían ser plausibles bajo la disitribución posterior predictiva.

La distribución posterior predictiva

La distribución posterior predictiva corresponde a la distribución de datos **no** observados ($\tilde{y}$), condicionados a los que **sí** observamos (y) y marginalizada sobre todos los parámetros del modelo, es decir:

$$p(\tilde{y}|y) = \int P(\tilde{y}|\theta)P(\theta|y)d\theta \tag{11}$$

Para hacer la evaluación de nuestro modelo vamos generar datos con esta distribución y compararlos con los datos observados. Si todo anda bien no deberíamos ver discrepancias significativas.

Pero cómo medimos las discrepancias? Una manera de hacerlo es definiendo *test quantities* sobre los aspectos del modelo que nos interesa que estén bien representados. Un test quantity $T(y, \theta)$ es una función escalar de los parámetros y los datos que opera sobre los datos observados y sobre los generados por el modelo. En el formato de probabilidad *clasica*, se suele usar el término *test statistic* porque opera **sólo** sobre los datos y **no** sobre los parámetros.

Tail area probabilities

El término *tail area probability*, o bien *p-value* se define clasicamente para un *test statistic* $T(y)$ como:

$$p_C = P(T(\tilde{y}) \geq T(y)|\theta) \tag{12}$$

donde la probabilidad de los datos se toma con θ fijo. Aquí usamos que la probabilidad de los $\tilde{y}$ dado los datos y y los parámetros θ es la misma que sólo condicionando en θ.

La idea de estos test estadísticos es representar en una ecalar una medida de diescrepancia entre las observaciones y que uno esperaría con el modelo suponiendo que los parámetros toman determinado valor.

Bayesian p-value En el caso bayesiano no consideramos que los parámetros toman un sólo valor, dado que la información que nos proveen los datos culmina en una distribución posterior de θs. Por lo que tiene sentido expandir la idea de *test statistic* a *text quantity* inncorporando la dependencia con θ. El p-value bayesiano está defindo como la probabilidad de que los datos generados por el modelo sean "más extremos"que los datos observados bajo los ojos del test:

$$P_B = P(T(\tilde{y}, \theta) \geq T(y, \theta)|y) \tag{13}$$

donde la probabilidad se evalua sobre la distribución posterior de θ y de los datos generados $\tilde{y}$.

En la práctica en general calculamos distribución posterior predictiva usando simulaciones. Si tenemos S simulaciones de la distribución posterior de θ, podemos tomar **sólo un** valor $\tilde{y}$ de $P(\tilde{y}|y)$ y así obtener S valores de $P(\tilde{y}, \theta|y)$. El checkeo lo hacemos comparando el test de los datos generados $T(\tilde{y}^s, \theta^s)$ con el de los observados $T(y, \theta^s)$. El valor de *p-value* será la proporción de veces que uno fue matyor que el otro ($T(\tilde{y}^s, \theta^s) \geq T(y, \theta^s)$).