Métodos Estadísticos Bayesianos con R

Distribuciones Iniciales y Funciones de Pérdida

Gibrán Peniche

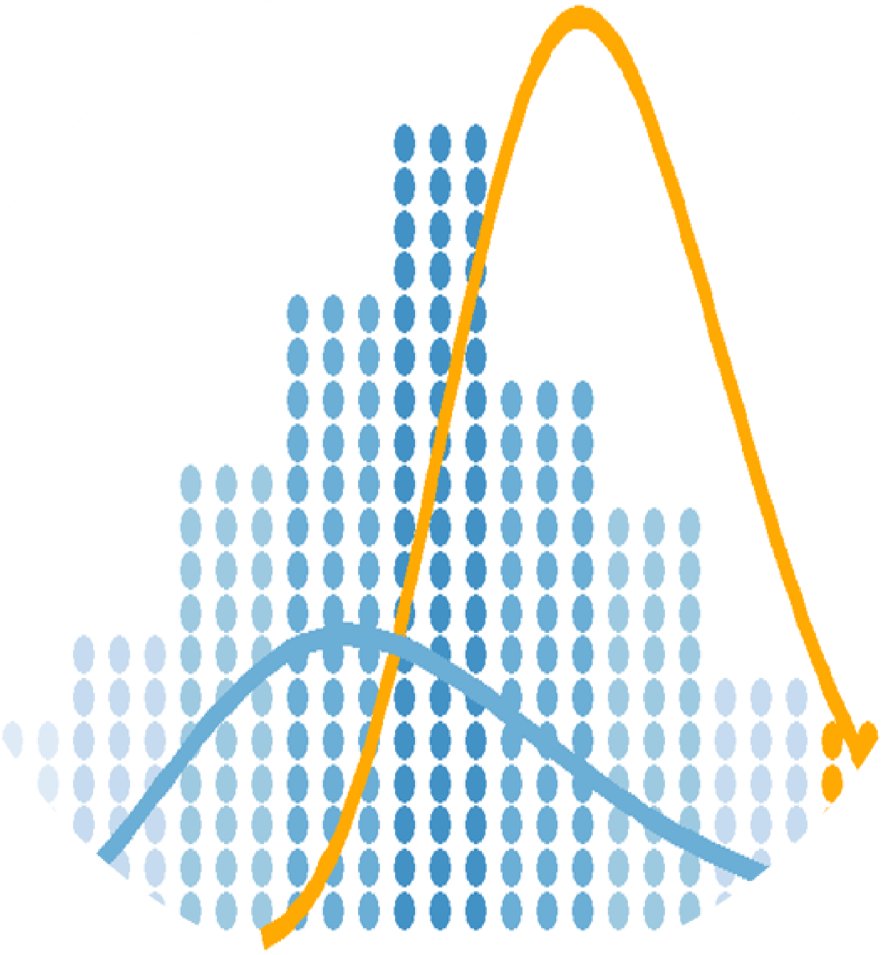
v. 0.0.1

2020-07-21

jgpeniche

PenicheGibran

jgpeniche@gmail.com



La sesión pasada...

Migramos de *maximizar* la **utilidad** u_0 a *minimizar* la **pérdida** $\mathcal{L}$, rescatando la noción de "distancia" del valor real θ

A partir de los Axiomas de Coherencia dedujimos un algoritmo de 3 pasos para nuestro problema de inferencia:

1. Definir una función de pérdida $\mathcal{L}$ y cuantificar la incertidumbre asociada a θ con $f(\theta)$ *a priori*

    1.1. En caso de existir información adicional (datos - verosimilitud) incorporarlos a través de Teorema de Bayes y obtener $P(\theta|X_{(n)})$

2. Minimzar la *pérdida esperada* $\int_{\Theta} \mathcal{L}(\theta, \hat{\theta}) P(\theta|X_{(\underline{n})})$ (en el caso **discreto**: $\sum_{\Theta} \mathcal{L}(\theta, \hat{\theta}) P(\theta|X_{(\underline{n})})$)

3. Escoger $d_{\hat{\theta}}$ que minimice dicha pérdida

Agenda

1. Entender las repercusiones de distintas funciones de pérdida

    1.1. Pérdida Cuadrática

    1.2. Pérdida Absoluta

    1.3. Pérdida (0,1)

2. Distribuciones iniciales

1

Funciones de Pérdida

Pérdida Cuadrática

Pérdida Cuadrática

Sea $\mathcal{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$

$$\arg\min_{\theta} \quad \int_{\Theta} (\hat{\theta} - \theta)^2 P(\theta | X_{(\underline{n})}) d\theta$$

Pero

$$\int_{\Theta} (\hat{\theta} - \theta)^2 P(\theta | X_{(\underline{n})}) d\theta = \hat{\theta}^2 \int_{\Theta} P(\theta | X_{(\underline{n})}) d\theta \quad - 2\hat{\theta} \int_{\Theta} \theta P(\theta | X_{(\underline{n})}) d\theta \quad + \int_{\Theta} \theta^2 P(\theta | X_{(\underline{n})}) d\theta$$

Además sabemos que $\int_{\Theta} P(\theta | X_{(\underline{n})}) d\theta = 1$

Pérdida Cuadrática

Tomando la derivada con respecto de $\hat{\theta}$ e igualando a 0

$$\frac{\partial}{\partial \theta} = 2\hat{\theta} - 2\int_{\Theta} \theta P(\theta | X_{(\underline{n})}) d\theta = 0$$

$\iff$

$$\hat{\theta} = \int_{\Theta} \theta P(\theta | X_{(\underline{n})}) d\theta = \mathbb{E}[\theta | X_{(\underline{n})}]$$

Pérdida Absoluta

Pérdida Absoluta

Sea $\mathcal{L}(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$

$$\arg\min_{\theta} \quad \int_{\Theta} |\hat{\theta} - \theta| P(\theta | X_{(\underline{n})}) d\theta$$

Pero

$$\int_{\Theta} |\hat{\theta} - \theta| P(\theta | X_{(\underline{n})}) d\theta = \int_{-\infty}^{\theta} (\hat{\theta} - \theta) P(\theta | X_{(\underline{n})}) d\theta \quad + \int_{\theta}^{\infty} (\theta - \hat{\theta}) P(\theta | X_{(\underline{n})}) d\theta$$

Pérdida Absoluta

Tomando la derivada con respecto a θ e igualando a 0

$$\int_{-\infty}^{\hat{\theta}} P(\theta|X_{(\underline{n})}) = \int_{\hat{\theta}}^{\infty} P(\theta|X_{(\underline{n})})$$

$\Longleftrightarrow$

$$\hat{\theta} = mediana$$

Pérdida (0,1)

Pérdida (0,1)

Sea

$$\mathcal{L} = \begin{cases} 1 & |\hat{\theta} - \theta| > \epsilon \\ 0 & |\hat{\theta} - \theta| \leq \epsilon \end{cases} = 1 - \delta(\theta \hat{-} \theta)$$

Donde *delta* denota la *delta de Dirc*

$\longrightarrow$

$$\int_{\Theta} (1 - \delta(\hat{\theta} - \theta)) P(\theta | X_{(\underline{n})}) d\theta = 1 - \int_{\Theta} \delta(\hat{\theta} - \theta) P(\theta | X_{(\underline{n})}) d\theta = 1 - P(\theta | X_{(\underline{n})})$$

El problema de minimizar la función objetivo es equivalente a maximizar la densidad $\therefore$

$$\hat{\theta} = moda$$

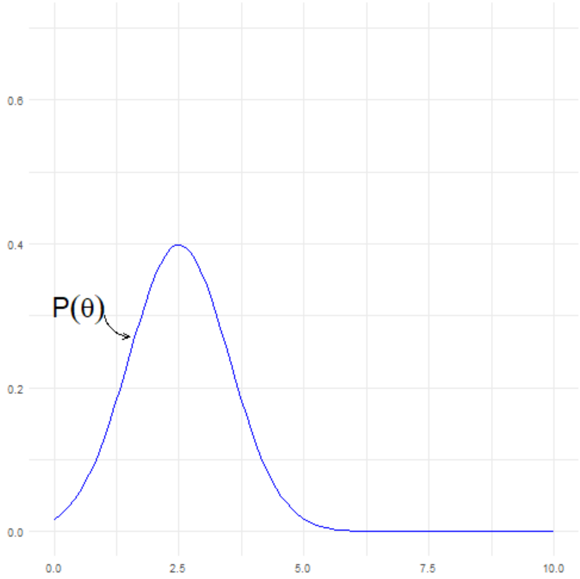
2

Distribuciones Iniciales

¿Cómo funciona el proceso de inferencia Bayesiano?

- $P(\theta|X_{(\underline{n})}) \propto \mathbb{L}(\theta|X_{(\underline{n})}) \cdot P(\theta)$ es una manera conciliar nuestra **incertidumbre** sobre el parámetro de interés y la información que aportan los *datos*

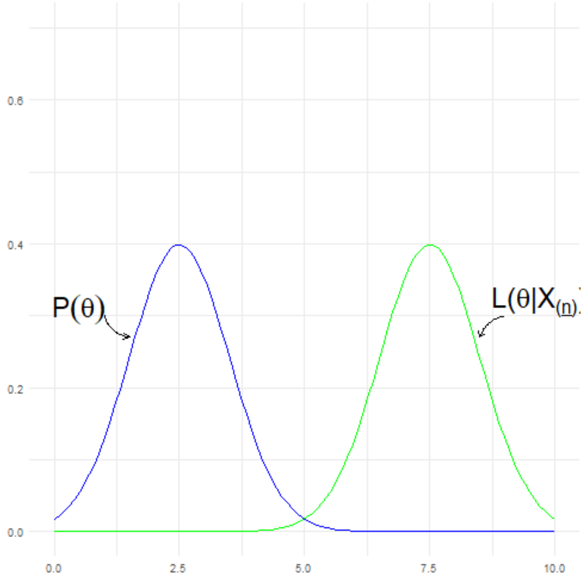
Gráficamente tenemos lo siguiente



¿Cómo funciona el proceso de inferencia Bayesiano?

- $P(\theta|X_{(\underline{n})}) \propto \mathbb{L}(\theta|X_{(\underline{n})}) \cdot P(\theta)$ es una manera conciliar nuestra **incertidumbre** sobre el parámetro de interés y la información que aportan los *datos*

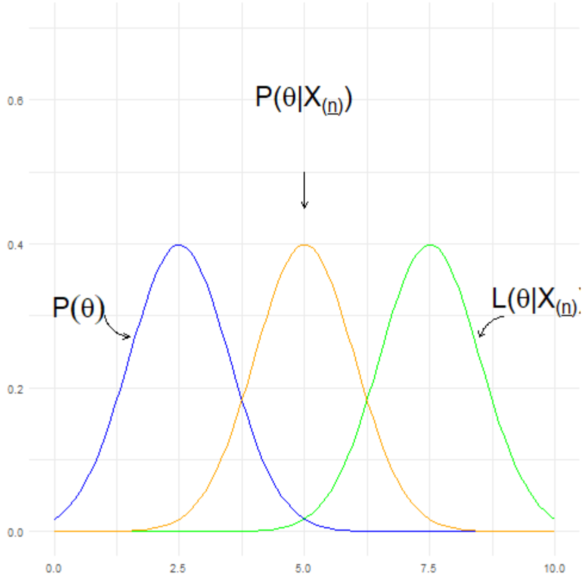
Gráficamente tenemos lo siguiente



¿Cómo funciona el proceso de inferencia Bayesiana?

- $P(\theta|X_{(\underline{n})}) \propto \mathbb{L}(\theta|X_{(\underline{n})}) \cdot P(\theta)$ es una manera conciliar nuestra **incertidumbre** sobre el parámetro de interés y la información que aportan los *datos*

Gráficamente tenemos lo siguiente



La pregunta es: ¿Cómo determinamos una $P(\theta)$ apropiada?

2

Distribuciones iniciales

Distribuciones iniciales

R: Depende...

- Recordemos que en el contexto del problema de inferencia $P(\theta)$ cuantifca nuestra incertidumbre alrededor del parámetro de interés

- En este sentido las preguntas que debe responder la elección de alguna distribución en particular debe responder al menos las siguientes preguntas:

    1. ¿ $P(\theta)$ es congruente con el espacio parametral Θ?

    2. ¿Está centrada alrededor de algún valor?

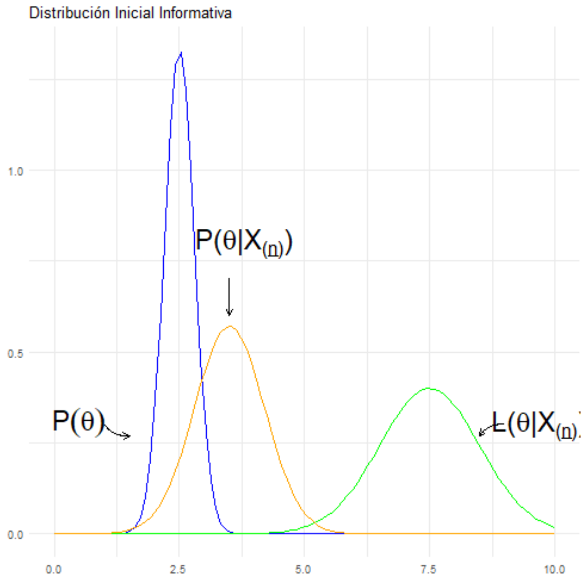
    3. ¿Es simétrica?

    4. ¿Qué tanta variabilidad presenta? Ó en otras palabras ¿Cuál es mi nivel de certidumbre medido en términos (p.e.) de la desviación estándar?

- Dependiendo de estas características en particular de la distribución inicial y su interacción con la verosimilutud la distribución posterior tendrá diferentes caracterísitcas

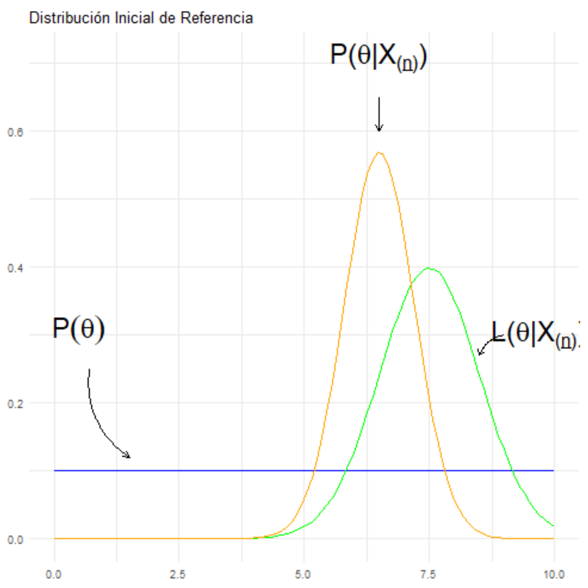
Distribuciones iniciales

- De acuerdo al grado de "certeza" del conocimiento *a priori* sobre el paramtero historicamente se les ha clasificado como distribuciones **informatvas** o **no informativas** (ó distribuciones **de referencia** en literatura más reciente)



Distribuciones iniciales

- De acuerdo al grado de "certeza" del conocimiento *a priori* sobre el paramtero historicamente se les ha clasificado como distribuciones **informatvas** o **no informativas** (ó distribuciones **de referencia** en literatura más reciente)



Distribuciones Iniciales de Referencia

- Uno de los motivos por los cuales Ronald A. Fisher, criticaba a la escuela Bayesiana era precisamente este elemento subjetivo intrínseco en la asignación de probabilidades *a priori* (Recordemos que la escuela frecuentista parte de postivismo de Augusto Comte por lo cual la fuente última de conocmiento es la experiencia)

- Además la asignación de probabilidades a través de la distribución uniforme es dificil de manipular al realizar la multiplcación de $\mathbb{L}(\theta|X_{(\underline{n})}) \cdot P(\theta)$

- Esto plantea un reto para la escuela bayesiana para encontrar un método de generar distribuciones de referencia. tales que se le diera "prioridad" a los datos y hacer la asignación *a priori* lo menos subjetiva posible

Pregunta: ¿Existe algún método para generar distribuciones de referencia que no sea la distribución uniforme?

Distribución Inicial de Jeffreys

- Sir Harold Jeffreys, fue un matemático, estadístico, geofísico y astrónomo británico es uno de los padres de la estadística Bayesiana

- Jeffreys concluyó que una opción posible para generar distribuciones de referencia para cualquier modelo es la siguiente:

$$P(\theta) \propto I(\theta)^{\frac{1}{2}}$$

- $I(\theta)$ denota la **Información de Fisher**, es decir

$$I(\theta) \propto -\mathbb{E}[\frac{\partial^2}{\partial\theta^2} ln f(x|\theta)]$$

Resiliencia de la Verosimilitud

- A pesar de que parece que el investigador puede forzar cierta distribución *a- posteriori* a través de la distribución *a-priori*, existe un umbral a partir del cual la interacción con la verosimilitud se queda fija ante distribucione sinciales 'locas'

- Al análisis (que debe acompañar a cualquier análisis Bayesiano) del umbral dónde la distribución *a posteriori* deja de cambiar ante distribuciones iniciales "extremas" se le llama **resilencia de la verosismilitud** (Haro _ Peniche, 2020)

- Esto implica que hay un "seguro" sobre que tanto puede el investigador incorporar información *a-priori* al experimento

Ejemplo Práctico

- Sea $x_i \sim Bernoulli(\theta)$ y $\theta \sim U(0, 1)$

Pregunta: ¿Cómo se distribuye $P(\theta|X_{(\underline{n})})$?

R: Sabemos que

$$P(\theta|X_{(\underline{n})}) \propto \mathbb{L}(\theta|X_{(\underline{n})})P(\theta)$$

Pero...

$$\mathbb{L}(\theta|X_{(\underline{n})}) = \prod_{i=1}^{n} f(x_i|\theta)$$

Como $x_i \sim Bernoulli(\theta) \implies$

$$\mathbb{L}(\theta|X_{(\underline{n})}) = \theta^{\sum_I x_i}(1-\theta)^{n-\sum_I x_i}$$

Ejemplo Práctico

$\Longrightarrow$

$$p(\theta|x(\underline{n})) \propto \theta^{\sum_I x_i}(1-\theta)^{n-\sum_{n-\sum I}} \frac{1}{\theta} 1_{(0,\theta)} 1_{(0,1)}$$

R: No hay respuesta cerrada

Ejemplo Práctico

- Sea $x_i \sim Bernoulli(\theta)$ y $\theta \sim \beta(2,2)$
- $$p(\theta|x(\underline{n})) \propto \theta^{\sum_I x_i}(1-\theta)^{n-\sum x_i}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$
- El operador $\propto$ implica que todo lo que no se realacione con el parámetro θ puede ser tratado como una constante
- $$p(\theta|x(\underline{n})) \propto \theta^{\sum_I x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$$
- $$P(\theta|X_{(\underline{n})}) \propto \beta(\theta|a=\sum_I x_i+\alpha, b=n-\sum x_i+\beta)$$

¿Qué elecciones de $P(\theta)$ y $x_i \sim f(x|\theta)$ resultan en una familia parametrica fácil de manipular?

¿Qué sigue?

1. Familias conjugadas